N U M E R I E K E   A N A L Y S E   II.

(Numerieke Lineaire Algebra)

Prof. Dr. A. van der Sluis.

1974

Mathematisch Instituut
der Rijksuniversiteit
te Utrecht.

Hoofdstuk V: Numerieke Lineaire Algebra

§ 16 Inleiding

Wij zullen onze aandacht concentreren rond de volgende problemen:

- Het oplossen van een stelsel van n lineaire vergelijkingen met n onbekenden.
- Het geven van een approximatieve oplossing voor een stelsel van n lineaire vergelijkingen met m onbekenden, $m < n$ (kleinste kwadraten problemen).
- Het bepalen van de eigenwaarden van een matrix. (Soms is men alleen geïnteresseerd in de absoluut grootste of absoluut kleinste eigenwaarde).

Regelmatig zullen wij ingaan op stabiliteitsaspecten bij de bovenstaande problemen en b.v. nagaan hoe stabiel eigenwaarden van matrices en oplossingen van stelsels lineaire vergelijkingen zijn t.o.v. afrondfouten.

Andere problemen die wij terloops zullen ontmoeten: inversie van matrices, berekening van determinanten, bepaling eigenvectoren.

§ 17 Lineaire ruimten en normen

(17.1) Zij X een n-dimensionale complexe (of reële) lineaire ruimte. De elementen van X zullen wij vaak noteren als kolomvectoren t.a.v. zekere basis $e_1, \ldots, e_n$, d.w.z.

$$x = \begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix} \quad \text{staat voor de vector } x = \sum_{i=1}^{n} x_i e_i$$

(17.2) Definitie. Een norm op X is een afbeelding $\| \ \| : X \to \mathbb{R}$, met de eigenschappen:

(i) $\|x\| = 0 \Leftrightarrow x = 0$

(ii) $\|\lambda x\| = |\lambda| \, \|x\|$ (voor alle $\lambda \in \mathbb{C}$ of $\mathbb{R}$)

(iii) $\|x + y\| \leq \|x\| + \|y\|$ (voor alle $x, y \in X$)

(17.3) Bij iedere norm op X hoort een limietbegrip als volgt: schrijf $\lim_{k\to\infty} y_k = y$ dan als voor alle $\varepsilon>0$ een N bestaat zodanig dat $\|y - y_k\| < \varepsilon$ voor alle $k > N$.

Zoals bekend bestaan voor elk tweetal normen $\|\ \|$ en $\|\ \|'$ op X positieve constanten γ_1 en γ_2 zodanig dat:

$$\forall_{x\in X}\ \gamma_1 \|x\| \leq \|x\|' \leq \gamma_2 \|x\|$$

Bijgevolg: als $\lim_{k\to\infty} y_k = y$ in zekere norm dan geldt dit voor elke andere norm op X ook. Alle normen op X geven dus aanleiding tot een zelfde limietbegrip.

(17.4) Bekende normen op X zijn de <u>Höldernormen</u> : $\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$ met $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, $p \geq 1$

(17.5) Voor $p = 1$: $\|x\|_1 = \sum_{i=1}^{n} |x_i|$ <u>lineaire norm</u>

(17.6) $p = 2$: $\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}$ <u>kwadratische of euklidische norm</u>

Via limietovergang ziet men dat voor $p = \infty$:

(17.7) $\|x\|_\infty = \max_{1\leq i\leq n} |x_i|$ <u>sup norm</u>

(17.8) <u>Opgave</u>. Toon aan

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\ \|x\|_2$$

$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\ \|x\|_\infty$$

Ga na dat de gelijktekens inderdaad kunnen optreden.

(17.9) <u>Definitie</u>. Een inproduct op X is een afbeelding $(.,.) : X\times X \to \mathbb{C}$ ($\mathbb{R}$) met de eigenschappen

(i) $(x,x) = 0 \Leftrightarrow x = 0$

(ii) $(x,y) = \overline{(y,x)}$

(iii) $(x,x) \geq 0$

(iv) $(\lambda x,y) = \lambda.\ (x,y)$ (dus $(x,\lambda y) = \bar{\lambda}(x,y)$)

(v) $(x+y,z) = (x,z) + (y,z)$

(17.10) Een eindig-dimensionale lineaire ruimte met inprodukt heet <u>euklidisch</u> in het reële, <u>unitair</u> in het complexe geval.

(17.11) Bij een gegeven basis $e_1, \ldots, e_n$ zullen wij $(x,y) = \sum_{i=1}^{n} x_i \cdot \bar{y}_i$ het standaard inprodukt noemen. Bij het standaard inprodukt is de basis vanzelf orthonormaal d.w.z. $(e_i, e_j) = \delta_{ij}$.

(17.12) Elk inprodukt induceert een norm $\|x\| = (x,x)^{\frac{1}{2}}$ op een unitaire ruimte X.

(17.13) Ga na dat de kwadratische norm geïnduceerd wordt door het standaard inprodukt.

§ 18 Lineaire operatoren en functionalen.

(18.1) Definitie Een lineaire afbeelding van X in zichzelf heet een lineaire operator (LO).

(18.2) Definitie Een lineaire afbeelding van $X \to \mathbb{C}$ heet een lineair functionaal (LF).

(18.3) Opgave. Ga na dat continuiteit van een afbeelding onafhankelijk is van de norm (vgl. (17.3)).

(18.4) Een norm $\| \;\|$ op X induceert normen voor een LO A resp. LF L, de zgn. geassocieerde normen die wij ook als $\| \;\|$ zullen schrijven:

(18.5) $$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\|$$

resp.

(18.6) $$\|L\| = \sup_{x \neq 0} \frac{|Lx|}{\|x\|} = \sup_{\|x\|=1} |Lx|$$

(18.7) Stelling. De geassocieerde norm bestaat voor elke LO en elk LF. Bovendien mag men in (18.5) en (18.6) "sup" vervangen door "max".

Bewijs. We zijn klaar indien we tonen dat een willekeurige LO A op n-dim. X continu is. De eenheidsbol ($\|x\|=1$) is immers compact, en een continue functie neemt op een compactum een maximum (voor $\|Ax\|$) aan. (Voor LF's gaat het analoog). Wegens (18.3) is het voldoende de continuiteit te bewijzen t.a.v. zekere norm op X. Wij kiezen hiervoor de sup norm (17.8), nadat we een basis in X gekozen hebben.

Stel dat A op de zojuist gekozen basis de matrix (a_{ij}) heeft.

Dan:

$$\|Ax\|_\infty = \max_{1\le i\le n} \left|\sum_{j=1}^{n} a_{ij}.x_j\right| \le \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| . \max_{1\le i\le n} |x_i| \le$$

$$\le c . \|x\|_\infty$$

Hieruit volgt direct de continuïteit.

(18.7a) Dus elke LO en LF is continu.

(18.8) Opgave. De LO's op X vormen een n^2-dimensionale lineaire ruimte, de LF's op X vormen een n-dimensionale lineaire ruimte. Ga dit na. Toon aan dat de geassocieerde normen normen zijn op de respectievelijke lineaire ruimten.

(18.9) Opmerking. Niet alle normen op de lineaire ruimte der LO's op X zijn geassocieerd aan een vectornorm.
Voorbeeld: stel de LO A heeft matrix (a_{ij}).
Zij $\|A\|_F = (\Sigma |a_{ij}|^2)^{\frac{1}{2}}$, de zgn. Frobeniusnorm.
Dan $\|I\|_F = \sqrt{n}$, terwijl voor elke geassocieerde norm $\|I\| = 1$.
In plaats van $\|A\|_F$ schrijft men ook wel $\|A\|_E$.

(18.10) Definitie. Een norm $\|.\|$ op de lineaire ruimte der LO's op X heet multiplicatief indien voor alle LO's A,B:

$$\|A.B\| \le \|A\| . \|B\| .$$

(18.11) Opgave. Geassocieerde normen zijn multiplicatief .
(Gebruik $\|Ax\| \le \|A\| . \|x\|$).

(18.12) Opgave. De Frobeniusnorm (zie (18.9)) is multiplicatief.

(18.13) Definitie. De abs. waarde van de absoluut grootste eigenwaarde van een (complexe) LO A heet de spectraalstraal $\rho(A)$ van A.

(18.14) Opgave. Voor een geassocieerde norm $\| \ \|$ geldt:
$\|A\| \ge \rho(A)$ (zie ook (20.6)).

(18.15) Voor iedere willekeurige nonsinguliere LO T geldt $\rho(A) = \rho(TAT^{-1})$.

§ 19 Matrices.

(19.1) Een lineaire afbeelding van n-dimensionale X in m-dimensionale Y kan bij vaste basis worden voorgesteld door een m x n - matrix. In het bijzonder kan een LO op X worden gerepresenteerd door een vierkante n x n - matrix, een LF door een 1 x n - matrix. Een vector is op te vatten als n x 1 - matrix.
We zullen in het vervolg LO's en LF's identificeren met hun matrix.

Zij A een matrix (a_{ij}), $i,j = 1,\dots,n$.

(19.2) Definitie. De geconjugeerde $\bar{A}$ van A is de matrix (b_{ij}) met $b_{ij} = \bar{a}_{ij}$

(19.3) Definitie. De gespiegelde of getransponeerde A^T van A is de matrix (c_{ij}) met $c_{ij} = a_{ji}$.

(19.4) Definitie. De geadjungeerde of complex-geconjugeerd gespiegelde A^* van A is de matrix (d_{ij}) met $d_{ij} = \bar{a}_{ji}$.

(19.5) Opmerking. Analoge definities gelden voor LF's en vectoren (opgevat als n x 1 - matrices). De gespiegelde van een LF is een vector en omgekeerd. Voor het natuurlijk inprodukt $(x,y) = \sum_{i=1}^{n} x_i\bar{y}_i$ schrijft men wel $(x,y) = y^*.x$.

(19.6) Opgave. Ga na dat $\bar{A}^T = A^*$, $A^{**} = A^{TT} = \bar{\bar{A}} = A$; $\overline{AB} = \bar{A}.\bar{B}$;
$(AB)^T = B^T A^T$; $(AB)^* = B^*A^*$, $(A^{-1})^* = (A^*)^{-1}$

(19.7) Opgave. Bij het natuurlijk inprodukt geldt $(Ax,y) = (x, A^*y)$.

(19.8) Definitie. Een matrix A heet hermiets indien $A = A^*$, unitair indien $A^{-1} = A^*$
(Voor een reële ruimte X heten zij symmetrisch resp. orthogonaal)

Hermietse en unitaire matrices zijn voorbeelden van zgn. normale matrices:

(19.9) Definitie. Een matrix A heet normaal indien $A^*A = AA^*$.

(19.10) Stelling. Voor elke normale matrix A bestaat een unitaire U zodat U*A U eendiagonaalmatrix is.
(Voor een bewijs zie bv. (20.14)).

(19.11) Opgave. Voor elke A is A*.A hermiets.

(19.12) Opgave. Stel er is een unitaire U zodat U*A U een diagonaalmatrix is. Toon aan dat A normaal is.

(19.13) Opgave. Ga na dat in het geval dat A hermiets is de diagonaalmatrix waartoe A unitair transformeerbaar is, reëel is. Geef het analogon van (19.12) voor hermietse A.

(19.14) Voor matrices in het algemeen bestaat de zgn. Jordan normaalvorm.

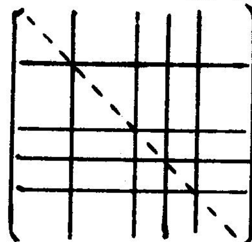
(19.15) Definitie. Een Jordan matrix is een vierkante matrix met op de diagonaal uitsluitend gelijke getallen, op de onmiddellijk rechts ervan gelegen nevendiagonaal (als het tenminste geen 1 x 1 - matrix is) enen, en verder nullen

$$\begin{pmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{pmatrix}$$

(19.16) Stelling. Bij elke n x n - matrix hoort een niet - singuliere matrix T, zodat $T^{-1}AT$ te schrijven is als aan elkaar geregen Jordanmatrices waarvan de diagonalen langs de diagonaal van A vallen, terwijl de overige matrixelementen nul zijn.

(19.17) Opmerkingen. We zullen deze stelling niet bewijzen. Zie voor een bewijs bv. Perlis p. 161, e.v. De Jordanvorm van A is, afgezien van de volgorde van de Jordanmatrices, eenduidig bepaald. De karakteristieke veeltermen van de Jordanmatrices zijn delers van de karakteristieke veelterm van A; men noemt ze de elementairdelers van A. Als alle elementairdelers van A lineair zijn en A dus op diagonaalgedaante gebracht kan worden, heeft A n onafhankelijke eigenvectoren, anders niet.

(19.18) Een Jordanvorm is een speciaal voorbeeld van een gepartitioneerde matrix i.e. een matrix die men door het aangeven van horizontale en evenzovele verticale scheidingslijnen heeft onderverdeeld in vakjes, waarbij de k^e horizontale en k^e verticale lijn elkaar snijden op de hoofddiagonaal:



(19.19) Opgave. Laten A en B gelijkgepartitioneerde matrices zijn. Duidt met A_{ij} en B_{ij} aan het blok op de i^e plaats van boven en de j^e plaats van links in A resp. B. Zij het produkt $C = AB$ weer op dezelfde wijze gepartitioneerd.
Dan is $C_{ij} = \Sigma A_{ik} B_{kj}$.

(19.20) Stelling. $\lim_{k\to\infty} A^k = 0 \Leftrightarrow \rho(A) < 1$

Bewijs.
$\Rightarrow$. Voor een geassocieerde norm geldt $\|A^k\| \geq \rho(A^k) = \rho(A)^k$.
$\Leftarrow$. Schrijf A als $T^{-1}BT$, B op Jordanvorm. Wegens $A^k = T^{-1}B^kT$ is dan te tonen $B^k \to 0$. Ga na dat het voldoende is te tonen dat voor elk Jordankastje C van B $C^k \to 0$.
Schrijf $C = \lambda I + D$, D een matrix met op en onder de hoofddiagonaal nullen, zodat $D^l = 0$ (l afmeting van C).
Voor $k > l$ geldt dus:

$$C^k = \lambda^k I + \binom{k}{1}\lambda^{k-1} D + \binom{k}{2}\lambda^{k-2} D^2 + \ldots + \binom{k}{l-1}\lambda^{k-l+1} D^{l-1}$$ en

dit gaat naar 0 voor $k \to \infty$.‖

(19.21) Opmerking. Als $\|A\| < 1$ in enige multiplicatieve norm, dan geldt trivialerwijze $\lim_{k\to\infty} A^k = 0$. Dat voor $\|A\| > 1$ toch $\lim_{k\to\infty} A^k = 0$ kan zijn blijkt uit $\begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix}$

(19.22) Opgave. $\|A\| \geq \rho(A)$ voor elke multiplicatieve norm.
N.B. Dit is een uitbreiding van (18.14).

(19.23) <u>Stelling.</u> $\sum_{k=0}^{\infty} A^k$ is convergent dan en slechts dan als $\lim_{k\to\infty} A^k = o$. De som is dan $(I - A)^{-1}$.

<u>Bewijs.</u> <u>Noodzakelijk</u> is triviaal. <u>Voldoende</u>
Alle eigenwaarden van A hebben modulus < 1 en dus is $\det(I - A) \neq 0$. $\Rightarrow (I - A)^{-1}$ bestaat.
Wegens $(I + A + A^2 + \ldots + A^k)(I - A) = I - A^{k+1}$ geldt
$(I + A + A^2 + \ldots A^k) = (1 - A^{k+1})(I - A)^{-1} = (I - A)^{-1} - A^{k+1}(I - A)^{-1}$
en het rechterlid convergeert naar $(I - A)^{-1}$ voor $k \to \infty$ ||.

(19.24) <u>Gevolg.</u> $I - A$ is niet-singulier als $\|A\| < 1$ voor een multiplicatieve norm.

(19.25) <u>Opgave.</u> Zij $\sum_{k=0}^{\infty} a_k z^k$ een machtreeks met convergentiestraal ρ.
Als $\rho(A) < \rho$ dan convergeert $\sum_{k=0}^{\infty} a_k A^k$. Toon dit aan.

§ 20 <u>Matrixnormen</u>

(20.1) Reeds in § 19 werden normen voor LO's ingevoerd. Dit zijn dus in feite normen voor matrices (vgl. (18.6)). Diverse van deze normen (met name de geassocieerde normen) kunnen op meer expliciete wijze gekarakteriseerd worden, hetgeen goed van pas zal blijken te komen. We zullen enkele identiteiten afleiden voor geassocieerde normen van Höldernormen ((17.4) e.v.).

(20.2) <u>Lemma.</u> $\|A\|_1 = \max_j \sum_i |a_{ij}|$ = maximale kolom absoluut som.

<u>Bewijs</u>: Ga na: $\|Ax\|_1 \leq \max_j \sum_i |a_{ij}| \cdot \|x\|_1$. Als het max. wordt aangenomen voor $j = k$, dan geldt het gelijkteken voor $x = e_k$, de k^e basisvector.

(20.3) <u>Lemma.</u> $\|A\|_\infty = \max_i \sum_j |a_{ij}|$ = maximale rij absoluut som.

<u>Bewijs.</u> Ga na $\|Ax\|_\infty \leq \max_i \sum_j |a_{ij}| \cdot \|x\|_\infty$. Als het max wordt aangenomen voor $i = k$, dan geldt het gelijkteken voor $x = (\mathrm{sgn}(a_{k1}), \ldots, \mathrm{sgn}(a_{kn}))^T$ als A reëel is. Hoe is het bij complexe A?

(20.4) Lemma. $\|A\|_2 = \max_{x\neq 0,\ y\neq 0} \frac{|(Ax,\ y)|}{\|x\|_2 . \|y\|_2}$

Bewijs.

(,) is het natuurlijk inprodukt dat dus de kwadratische norm induceert (zie (17.1)). Gebruik ongelijkheid van Schwarz, en bemerk dat het gelijkteken wordt aangenomen voor y = Ax.

(20.5) Ga na: $\|A\|_2 = \|A^*\|_2, \| U_1\ A\ U_2\ \|_2 = \|A\|_2$ voor alle unitaire $U_{1,2}$.

(20.6) Opgave. Ga na dat voor elke normale matrix A (dus i.h.b. voor hermietse en unitaire matrices) geldt: $\|A\|_2 = \rho(A)$.
(Gebruik de diagonaliseerbaarheid, (19.10)). Ga eveneens na dat $\|Ay\| \geq \min_i |\lambda_i(A)|\ \|y\|$, $\lambda_i(A)$ de eigenwaarden van A.

(20.7) Lemma. $\|A\|_2 = \rho(A^*A)^{\frac{1}{2}}$.

Bewijs.

$\|Ax\|_2 = (Ax,\ Ax)^{\frac{1}{2}} = (A^*A\ x,\ x)^{\frac{1}{2}}$; A^*A is hermiets dus heeft op geschikte basis diagonaalgedaante. Iets anders: Zij $U^*A^*AU = D$ (diagonaal) dan $\max_{\|x\|=1} (Ax,Ax) = \max_{\|Ux\|=1} (AUx,AUx) = \max_{\|x\|=1} (AUx,AUx) = \max_{\|x\|=1} (U^*A^*AUx,x) = \max_{\|x\|=1} (Dx,x)$

(20.8) Lemma. $\frac{\|A\|_1}{\|A\|_2}$, $\frac{\|A\|_\infty}{\|A\|_2}$, $\frac{\|A\|_1}{\|A\|_F}$, $\frac{\|A\|_\infty}{\|A\|_F}$ liggen alle tussen $\frac{1}{\sqrt{n}}$ en $\sqrt{n}$.

$\frac{1}{\sqrt{n}} \leq \frac{\|A\|_2}{\|A\|_F} \leq 1$.

De grenzen kunnen aangenomen worden.

Bewijs. Als opgave. Zie ook (20.15).

(Aanwijzing voor $\frac{\|A\|_2}{\|A\|_F}$: $\|A\|_F^2$ = spoor (A^*A) = som der eigenwaarden A^*A).

(20.9) Opgave. Zij $A = \begin{pmatrix} 0,9 & 0 \\ 0,3 & 0,8 \end{pmatrix}$

Laat zien dat $\|A\|_\infty, \|A\|_1, \|A\|_F$ alle > 1 zijn.

Toon convergentie van A^k naar 0 aan met (19.21) door een diagonaalmatrix D te vinden zodat $\|D^{-1}\ A\ D\| < 1$ in een door u zelf te kiezen norm.

(20.10) <u>Opgave</u>. Toon aan dat voor willekeurige multiplicatieve norm

$$\frac{\|A^{-1}\|}{1 + \|BA^{-1}\|} \leq \|(A+B)^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|BA^{-1}\|} \quad \text{mits } \|BA^{-1}\| < 1.$$

(20.11) <u>Opgave</u>. Zij A een n x n Jordanmatrix (n > 1) met diagonaalcoëff. λ.
Toon aan:

$$\frac{1}{|\lambda|+1} \leq \|A^{-1}\|_1 \leq \frac{1}{|\lambda|-1} \qquad (|\lambda| > 1)$$

$$\frac{\sqrt{n}}{|\lambda| + \sqrt{n(n-1)}} \leq \|A^{-1}\|_F \leq \frac{\sqrt{n}}{|\lambda| - \sqrt{n(n-1)}} \qquad (|\lambda| > \sqrt{n(n-1)})$$

Toets deze schattingen door de inverse van A expliciet te bepalen en de normen te berekenen.

(20.12) <u>Opgave</u> Toon aan dat voor een willekeurige multiplicatieve norm

$$\|(A+B)^{-1} - A^{-1}\| \leq \frac{\|BA^{-1}\|\,\|A^{-1}\|}{1 - \|BA^{-1}\|} \quad \text{als } \|BA^{-1}\| < 1.$$

<u>Hint</u> $(A+B)^{-1} - A^{-1} = -(A+B)BA^{-1}$

(20.13) <u>Hint</u> voor (20.10): $(A+B)^{-1} = A^{-1} - (A+B^{-1})BA^{-1}$, dus $\|(A+B)^{-1}\| \leq \ldots$ en $\geq \ldots$

(20.14) <u>Bewijs voor (19.10)</u>

a) Als men de kolommen van een matrix T orthonormaliseert met Gramschmidt krijgt men $T = UV$, U unitair, V bovendriehoeks

b) Bij elke A bestaat een T zodat $T^{-1}AT = W$, W bovendriehoeks, dus, met $T = UV$ volgens a) ook $V^{-1}U^{-1}AUV = W$ dus $U^{-1}AU = Z$, Z bovendriehoeks (belangrijke stelling van Schur: elke matrix is door unitaire transformaties op de driehoeksgedaante te brengen)

c) Er geldt $ZZ^* = Z^*Z$. Het (1,1) el. v.d. product matrix is voor ZZ^* $|z_{11}|^2 + |z_{12}|^2 + \ldots + |z_{1n}|^2$, voor Z^*Z is het $|z_{11}|^2$. $\Longrightarrow z_{12} = \ldots = z_{1n} = 0$. Etc.

(20.15) <u>Hints voor (20.8)</u>. Bepaal de 1, 2, ∞ en F norm van $\begin{pmatrix} 1 & 1 & \cdots & 1 \\ & & \Theta & \end{pmatrix}$ en van $\begin{pmatrix} 1 & \\ \vdots & \Theta \\ 1 & \end{pmatrix}$.

(20.16) <u>Belangrijke stelling</u> (<u>polaire ontbinding</u>) Voor een willekeurige m x n matrix A bestaan unitaire matrices U en V zodat $U^*AV = \Sigma$, Σ niet negatieve diagonaalmatrix.
<u>Bewijs</u> Neem V zo dat $V^*A^*AV = \Sigma^2$ is diagonaal. Interpretatie: de kolommen van AV zijn onderling orthogonaal, en de i-de kolom heeft norm $(\Sigma)_{ii}$. Dus is er een unitaire U zodat $AV = U\Sigma$ (immers $U\Sigma$ betekent: vermenigvuldig i^e kolom met $(\Sigma)_{ii}$)

(20.17) Men noemt de i^e kolom van V en de i^e rij van U resp. rechts- en links singuliere vector van A bij de singuliere waarde $(\Sigma)_{ii}$. Merk op $\|A\| = \max(\Sigma)_{ii}$, $\|A^{-1}\| = 1/\min(\Sigma)_{ii}$, mits bestaat.

Oplosmethoden voor stelsels van n lineaire vergelijkingen met n onbekenden.

§ 21 Gausseliminatie.

(21.1) Het stelsel heeft de vorm

$$\begin{matrix} a_{11}x_1 + \ldots\ldots + a_{1n}x_n = b_1 \\ \vdots \qquad\qquad \vdots \qquad \vdots \\ a_{n1}x_1 + \ldots\ldots + a_{nn}x_n = b_n \end{matrix}$$

en luidt dus in matrix notatie:

(21.2) $$A\ x = b$$

met $A = (a_{ij})$, de coëfficientenmatrix

$b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$, een gegeven kolomvector

$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, de gevraagde kolomvector.

(21.3) De eliminatie methode van Gauss komt in principe hierop neer: verdrijf x_1 uit de 2^e t/m n^e vergelijking door elk van deze vergelijkingen te verminderen met een geschikt veelvoud van de eerste vergelijking; verdrijf op analoge wijze x_2 uit de 3^e t/m n^e vergelijking mbv. de 2^e vergelijking etc. Tenslotte heeft men het stelsel teruggebracht tot een stelsel met driehoeksmatrix, waaruit x_n triviaal oplosbaar is, substitutie hiervan in de (n-1)-ste vergelijking levert onmiddellijk x_{n-1} etc.

(21.4) Vanzelfsprekend loopt Gauss-eliminatie mis als $a_{11} = 0$. Algemener, als we bij Gauss-eliminatie met $A^{(k)}$ aangeven de matrix van het stelsel dat ontstaat na eliminatie van x_{k-1} (dus $A^{(1)} = A$) dan loopt het proces mis als $a_{kk}^{(k)} = 0$. Wanneer echter het oorspronkelijk stelsel onafhankelijk is kan niet gelden dat $a_{kk}^{(k)}$ t/m $a_{nk}^{(k)}$ alle 0 zijn. Men bedenke hierbij dat $A^{(k)}$ van de vorm

$$\begin{pmatrix} * & & & & \\ & \ddots & & * & \\ 0 & & * & & \\ & & \vdots & & \vdots \\ 0 & & * & & * \end{pmatrix}$$

is en dat $\det A = \det A^{(k)}$.

(21.5) Derhalve: Als A niet singulier is kan men tijdens het eliminatie proces de vergelijkingen steeds zo omnummeren dat $a_{kk}^{(k)} \neq 0$.

(21.6) Men komt zo tot de volgende algoritme: kijk, alvorens x_k te elimineren of $a_{kk}^{(k)} \neq 0$ is, en zo niet, zoek een $a_{ik}^{(k)}$, $i = k+1,\ldots,n$, die niet 0 is, en verwissel de i-de en k-de vergelijking.

(21.7) De gebruikte elementen $a_{kk}^{(k)}$ noemt men de **pivots** (=spillen) en het zoeken van een $a_{ik}^{(k)}$ die niet 0 is plus het vervolgens verwisselen noemt men pivoting.

(21.8) Opgave. Ga na: $A^{(2)} = a_{ij}^{(2)}$ met

$$a_{ij}^{(2)} = a_{ij}^{(1)} \quad \text{als } i = 1$$
$$= a_{ij}^{(1)} - \frac{a_{i1}^{(1)} \cdot a_{1j}^{(1)}}{a_{11}^{(1)}} \quad \text{als } i > 1$$

(21.9) Opgave. Laat $A^{(k)} = \left(a_{ij}^{(k)}\right)$, $a_{kk}^{(k)} \neq 0$.

Ga na: $A^{(k+1)} = \left(a_{ij}^{(k+1)}\right)$, met

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} \quad \text{als } i \leq k.$$
$$= a_{ij}^{(k)} - \frac{a_{ik}^{(k)} \cdot a_{kj}^{(k)}}{a_{kk}^{(k)}} \quad \text{als } i > k.$$

(21.10) Opgave. Laat $A^{(k)} = (a_{ij}^{(k)})$, $a_{kk}^{(k)} \neq 0$.

Ga na: $A^{(k+1)} = L_k \, A^{(k)}$

met

$$L_k = \begin{pmatrix} 1 & & & & \theta \\ & \ddots & & & \\ \theta & & 1 & & \theta \\ & & -m_{k+1\,k} & \ddots & \\ & & \vdots & & \\ \theta & & -m_{nk} & \theta & 1 \end{pmatrix}$$

waarin

$$m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \quad (i > k)$$

(21.11) Opgave. Stel dat tijdens het eliminatieproces steeds $a_{kk}^{(k)} \neq 0$. U weet dat dit eventueel na vernummeren inderdaad zo is mits $\det(A) \neq 0$.

Ga na: $\det A = a_{11}^{(1)} \cdot a_{22}^{(2)} \cdots\cdots\cdots a_{nn}^{(n)}$.

§ 22. Practische uitvoering van Gausseliminatie. De hoeveelheid rekenwerk.

(22.1) Bij Gauss-eliminatie wordt het stelsel $Ax=b$ stapsgewijs getransformeerd naar stelsels $A^{(k)}x=b^{(k)}$ $k=1,\dots,n$, waarbij men voor $k=n$ op een driehoeksstelsel uitkomt.
In de praktijk voert men gewoonlijk eerst alle transformaties van de matrix A uit en berekent pas later de corresponderende transformaties van het rechterlid b. (we zullen straks zien waarom).

(22.2) Stel dat de elementen van matrix A in de rekenmachine zijn opgeslagen in een vierkant array C.
In feite voert men dan alle operaties uit op dit array, waarvan we de elementen aanduiden zullen met: c_{ij}.
In het begin geldt dus:

$$c_{ij}=a_{ij}^{(1)} \quad ,$$

na het elimineren van x_1:

$$c_{ij}=a_{ij}^{(2)}$$

etc.
We houden ons voorlopig niet bezig met pivoten, dus nemen aan dat steeds $a_{kk}^{(k)}\neq 0$.

(22.3) Nu is het niet erg interessant in de eerste kolom van C, voorzover onder de diagonaal gelegen, al die nullen op te schrijven die door het elimineren van x_1 ontstaan.
We weten toch wel dat op die plaatsen van $A^{(2)}$ nullen staan, en kunnen die plaatsen van C wel voor iets anders gebruiken.
Men plaatst nu in c_{i1} $(i=2,\dots,n)$ het getal

$$m_{i1}=\frac{a_{i1}^{(1)}}{a_{11}^{(1)}}$$

i.e. de factor waarmee men de eerste rij moest vermenigvuldigen alvorens deze van de i^e rij af te trekken.
Evenzo plaatst men in de benedendiagonaalelementen van de tweede kolom van C de getallen

$$m_{i2} = a_{i2}^{(2)} / a_{22}^{(2)} \quad (i=3,\dots,n) \; ,$$

dus de factoren waarmee men de 2^e rij van $A^{(2)}$ moet vermenigvuldigen om x_2 uit de overige vergelijkingen te elimineren.
Etc.

(22.4) Na afloop ziet C eruit als in de nevenstaande figuur, waarbij de u's juist de elementen van de uiteindelijke driehoeksmatrix $U=A^{(n)}$ zijn, en de m's de elementen m_{ij} voorstellen. Met de getallen m_{ij} kunnen we nu het rechterlid van het stelsel gaan transformeren (vgl. (22.1))

$$\begin{array}{cccccc} u & u & u & u & u & u \\ m & u & u & u & u & u \\ m & m & u & u & u & u \\ m & m & m & u & u & u \\ m & m & m & m & u & u \\ m & m & m & m & m & u \end{array}$$

$$b^{(1)} = b$$

$$b_i^{(2)} = b_i^{(1)} - m_{i1} b_1^{(1)} \qquad i \geq 2$$

$$b_i^{(3)} = b_i^{(2)} \quad m_{i2} b_2^{(2)} \qquad i \geq 3$$

$$\vdots$$

(22.5) Het is nu wel duidelijk waarom we eerst alle transformaties op A en daarna pas die op b hebben uitgevoerd(vgl. (22.1)):

als er meerdere stelsels zijn met dezelfde A, maar met verschillende b, dan hoeven we de operaties op A slechts éénmaal uit te voeren.

(22.6) In het algemeen zal natuurlijk wel gepivot moeten worden. Stel b.v. dat voor k=3 voor het eerst $a_{kk}^{(k)} = 0$, dus $a_{33}^{(3)}=0$ en zij $a_{53}^{(3)} \neq 0$.
Dan willen we in $A^{(3)}$ de 3^e en de 5^e rij verwisselen, maar de vraag is wat we dan met de m's in de 1^e en 2^e kolom van C moeten doen.
Ga na dat als we in C de <u>gehele</u> 3^e en 5^e rij verwisselen (dus óók de m's op die rijen) we precies dezelfde matrix C krijgen als wanneer we in de oorspronkelijke matrix A de 3^e en 5^e rij verwisseld hadden.
Als later nog eens $a_{kk}^{(k)}$ nul blijkt te zijn, maar $a_{rk}^{(k)} \neq 0$ verwissel dan weer de gehele k^e en r^e rij van C, enzovoorts.

(22.7) <u>Opgave</u>. Verifieer de bovenstaande beweringen.

(22.8) De na Gauss-eliminatie met pivoting resulterende matrix C past aldus bij een rij-gepermuteerde matrix A, en we moeten het rechterlid dus aan dezelfde permutatie onderwerpen.

(22.9) Er geldt nog een opmerkelijke eigenschap voor de uiteindelijk resulterende elementen van C.
Zij U de bovendriehoeksmatrix (i.e. een matrix met onder de hoofddiagonaal nullen) die op en boven de hoofddiagonaal de u's uit figuur (22.4)heeft, dus $U=A^{(n)}$.
Zij L de benedendriehoeksmatrix met op de diagonaal louter enen en onder de diagonaal het resterende gedeelte van C, dus de m's uit figuur (22.4).

(22.10) <u>Stelling</u>. Indien bij Gauss-eliminatie geen rijverwisselingen plaatsgevonden hebben geldt:

$$A = L.\ U.$$

<u>Bewijs</u>. In (21.10) bleek : $A^{(k+1)} = (I-M_k)A^{(k)}$, met M_k de matrix die bijna geheel uit nullen bestaat en alleen in de k^e kolom onder de diagonaal op de plaats (i,k) $(i > k)$ het getal m_{ik} heeft.

Dan: $U = A^{(n)} = (I-M_{n-1}) \ldots\ldots (I-M_1)\ A^{(1)}$

Wegens

$$(I-M_k)^{-1} = I + M_k \qquad \text{(Ga na)}$$

geldt

$$A = A^{(1)} = (I+M_1) \ldots\ldots (I+M_{n-1}).A^{(n)}$$

Ga na dat juist $L=(I+M_1) \ldots (I+M_{n-1})$.

(22.11) Het Gauss-eliminatieproces komt dus neer op decompositie van de na het eventueel rijverwisselen (pivoting!) verkregen matrix A als produkt van een beneden- en een bovendriehoeksmatrix.
Men spreekt van een <u>L-U-decompositie</u>.

(22.12) Stel A is nonsingulier en heeft een L-U-decompositie.
Toon aan dat voor 2 verschillende L-U-decomposities $A=L\ U$ en $A' = L'U'$ een diagonaalmatrix D bestaat zodat $L=L'D$ en $U' = DU$.
Indien van L of U de diagonaal voorgeschreven is, dan is de decompositie derhalve uniek.

(22.13) We bezien nog de hoeveelheid rekenwerk die nodig is voor het oplossen van een lineair stelsel. Als eenheid nemen we de <u>accumulatieve vermenigvuldiging</u> (AV), i.e. een vermenigvuldiging gevolgd door een optelling. Ook een deling beschouwen we voor wat de hoeveelheid rekenwerk betreft als 1 AV.

(22.14) Opgave. Verifieer de volgende beweringen
Berekening van een inprodukt in $\mathbb{R}_n$ kost n AV
Matrix x vector kost n^2 AV
Matrix x matrix kost n^3 AV

(22.15) Het berekenen van elke m (zie(22.3)) kost 1 deling ≡ 1 AV., dus in totaal $\frac{1}{2}n\ (n-1)$ AV.
De eerste k rijen van $A^{(k)}$ en $A^{(k+1)}$ zijn gelijk, het bepalen van de overige elementen van $A^{(k+1)}$ uit die van $A^{(k)}$ als de m eenmaal bekend is kost 1 AV per element.
Voor $A^{(2)}$ wordt dit dus $(n-1)^2$ AV, voor $A^{(3)}$ $(n-2)^2$ AV, etc.
In totaal vergt bepaling van $A^{(n)}$ dus

$$\sum_{k=1}^{n-1} (n-k)^2 = \sum_{k=1}^{n-1} k^2 = \frac{1}{6}\, n(n-1)(n-2) \text{ AV}$$

en de bepaling van het uiteindelijke array C
$\frac{1}{2}n(n-1) + \frac{1}{6}\, n(n-1)(n-2) = \frac{1}{3}\, n(n^2-1)$ AV.
De in (22.4) genoemde operaties op het rechterlid kosten

$$(n-1) + (n-2) + \ldots + 1 = \tfrac{1}{2}n\ (n-1) \text{ AV.}$$

Het oplossen van het driehoeksstelsel $A^{(n)}x = b^{(n)}$ vergt nog eens $\frac{1}{2}n(n+1)$AV.

(22.16) Totaal vergt Gauss-eliminatie: $\frac{1}{3}n^3 + n^2 - \frac{1}{3}n$ AV.
d.i. $\approx \frac{1}{3}n^3$ voor n groot, hetgeen verrassend weinig is in vergelijking met een zo eenvoudige bewerking als matrix vermenigvuldiging.

(22.17) Heeft men de uiteindelijke C eenmaal verkregen dan kost oplossen per rechterlid slechts $\frac{1}{2}n(n-1) + \frac{1}{2}n(n+1) = n^2$ AV.

Dit is dezelfde hoeveelheid werk als vereist voor matrix x vector, zodat het ook bij vele rechterleden niet loont eerst de matrix te inverteren. Immers, ook als men over A^{-1} beschikt kost het berekenen van A^{-1} b nog n^2 AV.

(22.18) Opgave. Door in Ax= b voor b achtereenvolgens de basis-vectoren $e_1,\ldots, e_n$ te kiezen, kan men de inverse van A berekenen. Ga na dat bij gebruik van Gauss-eliminatie dit kan geschieden in : n^3 AV.

(22.19) Opgave. Ga voor verschillende waarden van n (bv. 10 en 100) na wat de getallen betekenen bij een rekensnelheid van bv. 50 microsec/AV.

(22.20) Opgave. De hoeveelheid rekenwerk bij Gauss-eliminatie steekt heel gunstig af bij de hoeveelheid werk vereist voor de regel van Cramer wanneer men de daarin voorkomende determinanten m.b.v. de permutatieregel berekent ((n+1)!(n-1) vermenigvuldigingen alleen).

§ 23. Varianten van Gauss-eliminatie: Doolittle, Crout, Choleski.

(23.1) Er zijn diverse andere methoden die zich slechts van Gauss-eliminatie onderscheiden door de volgorde waarin de bewerkingen worden uitgevoerd, hetgeen voor computer-gebruik zekere organisatorische voordelen kan bieden.

(23.2) Een heel bekende rekenvolgorde is afkomstig van Doolittle. Stel dat in nevenstaande figuur de aangegeven elementen u en m reeds overeenstemmen met die van het array C uit(22.4), zoals ze zijn na het elimineren van x_1 en x_2. Laten de plusjes nog de oorspronkelijke elementen van A zijn (afgezien van eventuele rijverwisselingen).

u	u	u	u	u	u
m	u	u	u	u	u
m	m	+	+	+	+
m	m	+	+	+	+
m	m	+	+	+	+
m	m	+	+	+	+

De u's zijn dus elementen van $A^{(3)}$, de ontbrekende elementen van $A^{(3)}$ zijn zonder meer te berekenen: immers

$a_{ij}^{(2)} = a_{ij}^{(1)} - c_{i1} \cdot c_{1j}$ en $a_{ij}^{(3)} = a_{ij}^{(2)} - c_{i2}c_{2j}$ (vgl.(21.9),(22.2))

dus $a_{ij}^{(3)} = a_{ij}^{(1)} - c_{i1}c_{1j} - c_{i2}c_{2j}$.

Maar we berekenen eerst alleen $a_{33}^{(3)}$ t/m $a_{63}^{(3)}$.

(23.3) Uit die elementen $a_{33}^{(3)}$ t/m $a_{63}^{(3)}$ kunnen we immers de pivot zoeken en aansluitend eventueel rijen verwisselen (vgl. (22.6)). We berekenen dan de resterende elementen van de 3^e rij van $A^{(3)}$, en dat zijn meteen elementen van $A^{(4)}$. Tenslotte delen we c_{43} t/m c_{63} door c_{33} en hebben dan de nieuwe m's (vgl. (22.3)).

Dus: uit de in fig. (23.2) aangegeven elementen van C zoals ze zijn na het elimineren van x_1 en x_2 kunnen we de in nevenstaande figuur aangegeven elementen van C zoals deze zijn na het elimineren van x_1 t/m x_3 bepalen.

u	u	u	u	u	u
m	u	u	u	u	u
m	m	u	u	u	u
m	m	m	+	+	+
m	m	m	+	+	+
m	m	m	+	+	+

(23.4) Het zal duidelijk zijn hoe men het Doolittle-proces van begin tot eind doorvoert.
Het is numeriek equivalent met Gauss-eliminatie volgens § 22, d.w.z. de afrondfouten in de berekende coëfficienten van L en U zijn bij beide methoden gelijk (mits men bij beide methoden steeds dezelfde pivots kiest).

(23.5) Het belang van Doolittle's methode is drieledig:

- Men hoeft niet telkens tussenresultaten in C op te bergen. Als men afziet van rijverwisselingen komt op elke plaats van C zelfs meteen het uiteindelijke element terecht. Dit is overigens maar een gering voordeel bij gebruik van een computer. Het was belangrijk bij het werken met tafelrekenmachines
- De coëfficienten $a_{ij}^{(k)}$ berekent men bij Doolittle in feite door van $a_{ij}^{(1)}$ een inprodukt af te trekken. (vgl. (23.2)). Datzelfde gebeurt bij Gauss kennelijk ook, omdat daar dezelfde bewerkingen worden uitgevoerd, maar dat inprodukt wordt bij Gauss stapsgewijs opgebouwd (per stap Gauss 1 term erbij).

Nu is bij de meeste programmeertalen de berekening van een expliciet uitgeprogrammeerd inprodukt (en dat heeft men bij Gauss), nogal tijdrovend.
Vandaar dat men vaak speciale snelle routines heeft voor het berekenen van inprodukten, en die kan men bij Doolittle gebruiken. Dit spaart gewoonlijk een factor op de rekentijd.

- Als men bij Gauss in dubbele precisie wenst te rekenen, en de uiteindelijke matrix $A^{(n)}$ weer op enkele precisie wil afronden, heeft men voor de hele matrix dubbele precisie geheugenplaatsen nodig is, i.e. 2x zoveel ruimte als bij enkele precisie. Bij Doolittle kan men het accumuleren van de inprodukten in dubbele precisie uitvoeren en de verkregen elementen van $A^{(n)}$ na afronding in enkele precisie in het geheugen opbergen.

(23.6) Als men afziet van het pivot zoeken is Doolittle in feite een directe manier om de LU-decompositie van A te bepalen (zie (22.11)).
Men komt dan nl. op precies dezelfde inproduktformules als in (23.2) : $a_{ij} := \Sigma\ l_{ik}\ u_{kj}$.

(23.7) Bij Doolittle had men de eis dat L een diagonaal van louter enen had.
Een soortgelijk proces is dat van <u>Crout.</u>
Bij Crout wordt een decompositie L'U' bepaald, waarin L' een benedendriehoeksmatrix is, U' een bovendriehoeksmatrix met louter enen op de hoofddiagonaal. Ga zelf na hoe de coëfficienten van L' en U' in één rekengang bepaald kunnen worden en hoe het pivotten dient te geschieden.

(23.8) Voor symmetrische positief definiete matrices, i.e. symmetrische matrices met de eigenschap dat $(Ax,x) > 0$ voor $x \neq 0$, gebruikt men veelal de methode van <u>Choleski.</u>
Deze berust op de volgende stelling:

(23.9) <u>Stelling</u>. Zij A symmetrisch, positief definiet.
Dan bestaat een benedendriehoeksmatrix G zodat

$$A = G.G^T.$$

Bewijs. Met inductie kan men tonen dat A zonder rijverwisselen een L U-decompositie heeft waarbij de hoofddiagonaal van U uit louter positieve elementen bestaat.
Er bestaat een diagonaalmatrix D(die dan noodz. positieve diagonaalelementen heeft) en een bovendriehoeksmatrix U' met louter enen op de hoofddiagonaal zodat U = D U'.
Dan A = L D U'.
Ga na dat L, D en U' eenduidigbepaald zijn (zie (22.12)).
Symmetrie impliceert : $L^T = U'$, zodat

$$A = L\,D\,L^T = L\,D^{\frac{1}{2}}D^{\frac{1}{2}}L^T = L\,D^{\frac{1}{2}}.\,(L\,D^{\frac{1}{2}})^T.$$

(23.10) Opmerking. In bovenstaand bewijs is $D^{\frac{1}{2}}$ gedefinieerd door $D^{\frac{1}{2}}.\,D^{\frac{1}{2}} = D$. Ga na dat $D^{\frac{1}{2}}$ een diagonaalmatrix is met reële coëfficienten.

(23.11) Opgave. Ga na dat de decompositie in (23.9) uniek is, op het teken der kolommen na. (vgl.(22.12)).

(23.12) Bij Choleski bepaalt men $G = (g_{ij})$ op de volgende wijze:

bepaal g_{11}

bepaal g_{i1} $(i > 1)$

bepaal g_{22}

bepaal g_{i2} $(i > 2)$

etc.

(23.13) Merk op dat het vanwege de symmetrie van A overbodig is het boven- (of beneden-) diagonaal gedeelte van A in het computergeheugen op te slaan. De elementen van de j^e kolom van G kan men schrijven op de plaatsen van de elementen a_{ij} $(i \geq j)$ van de j^e kolom van A (Ga na).
Men ziet dat aldus de hoeveelheid werk en geheugenruimte nagenoeg gehalveerd wordt (in vergelijking tot b.v. Doolittle en Crout.)
Een tweede voordeel is dat men bij deze methode niet behoeft te pivotten.

§ 24 Perturbatie van stelsels lineaire vergelijkingen.

(24.6), (24.7), (24.8) en het bewijs van (24.9) kunnen door niet hoofdvak-studenten wiskunde worden overgeslagen.

§ 24.1. We zullen later zien, dat de afrondfouten tijdens het oplossen van het stelsel vergelijkingen $Ax = b$ gemaakt, geïnterpreteerd kunnen worden als een geringe verstoring van de matrix A. D.w.z. door berekening verkrijgt men i.p.v. de ware oplossing x een vector $\tilde{x}$ waarvoor geldt $(A + \Delta A)\tilde{x} = b$.
Ook gebeurt het vaak dat de elementen van A en b zelf door berekening of meting verkregen zijn en dus evt. (licht) verstoord.
We vragen ons daarom af wat het effect van perturbaties δA en δb is op de oplossing van het stelsel $Ax = b$.
Zij dus

(24.1) $(A + \delta A)\tilde{x} = b + \delta b$.
met $A + \delta A$ non-singulier.
We zijn geïnteresseerd in $\tilde{x} - x = \Delta x$. Er geldt:
$\Delta x = (I + A^{-1}\delta A)^{-1}A^{-1}(\delta b - \delta A x)$. Als nu $\|A^{-1}\delta A\| \ll 1$ in een relevante geassocieerdenorm dan geeft

(24.2) $$\delta x = A^{-1}(\delta b - \delta A x)$$

een goede benadering voor Δx. (ga na) Deze δx zullen we voortaan beschouwen.

(24.3) Opgave. Ga na, dat voor de differentialen dA, dx, db geldt:
$$dx = A^{-1}(db - dAx).$$
Dit betekent, dat δx een eerste orde benadering van Δx is.

Uit (24.2) volgt:

(24.4) $$\|\delta x\| \leq \|A^{-1}\| \; \|\delta A\| \; \|x\| + \|A^{-1}\| \; \|\delta b\|.$$

We vragen ons af of het gelijkteken in (24.4) kan optreden, m.a.w. hoe grof ongelijkheid (24.4) is. Daartoe introduceren we het begrip maximaliserende vector.

(24.5) Definitie. We noemen $x \in X$ voor een gegeven vector norm maximaliserende vector van een LO A of een LF L als voor de geassocieerdenorm geldt $\|Ax\| = \|A\| \; \|x\|$ resp. $|Lx| = \|L\| \; \|x\|$.

(24.6) Stelling. (Hahn - Banach) Ieder reëel LF F gedefiniëerd op een lineaire deelruimte X_0 van een reële lineaire ruimte X is met behoud van norm uit te breiden tot een LF op X.
Bewijs. Ieder boek over functionaalanalyse.

(24.7) Stelling. Voor een gegeven $x \in X$, $x \neq 0$, en gegeven norm op X is er een LF L op X waarvoor x maximaliserende vector is.
Bewijs. Zij V de verzameling van de veelvouden van x. Definieer op V een LF L door $Ly = \alpha\|x\|$ als $y = \alpha x$. Dan is $|Ly| = \|y\|$ voor alle $y \in V$, dus $\|L\| = 1$ en elke $y \in V$ is maximaliserende vector. Na uitbreiding van L volgens Hahn - Banach geldt nog steeds $\|L\| = 1$, en dus blijft x een maximaliserende vector.

(24.8) Gevolg. Voor $x, c \in X$, $x \neq 0$, en een gegeven norm op X is er een LO B op X waarvoor x maximaliserende vector is en waarvoor $Bx = c$.
Dit ziet men in door B te definiëren als $By = \frac{(Ly)c}{\|x\|}$, L het in (24.7) bedoelde LF.

((24.9) Stelling. In ongelijkheid (24.4) kan het gelijkteken optreden voor gegeven $\|\delta A\|$ en $\|\delta b\|$ en voor een gegeven norm op X en alle A en b.
Bewijs: Zij m een maximaliserende vector van A^{-1}, $\|m\| = 1$. Kies $\delta b = \|\delta b\|.m$. Op grond van (24.8) is er een δA van gegeven norm $\|\delta A\|$ zo, dat x maximaliserende vector van δA is en $\delta Ax = -\|\delta A\|\,\|x\|m$.

§ 24.2 A posteriori schatting voor de perturbaties van het oplosproces.

Vooraf enkele definitees :

(24.10) Definitie. $A = (a_{i,j})$ en $B = (b_{i,j})$.
We definiëren $A > B$ indien $a_{i,j} > b_{i,j}$ voor alle i en j.

(24.11) Definitie. a,b vectoren.
We definiëren $a > b$ indien $a_i > b_i$ voor alle i.

(24.12) Notatie. Zij A de matrix $(a_{i,j})$
met $|A|$ geven we aan de matrix $(|a_{i,j}|)$. Zij b de vector (b_j); dan is $|b|$ de vector $(|b_j|)$.

(24.13) We beschouwen **een gegeven stelsel vergelijkingen Ax=b. Stel nu,** dat de $a_{i,j}$ met onzekerheden Δa_{ij} en de b_i met onzekerheden Δb_j belast zijn. Men zou dan alle stelsel $\tilde{A}\tilde{x} = \tilde{b}$ met $|\tilde{A}-A| \leq \Delta A$ en $|\tilde{b}-b| \leq \Delta b$ mogelijke stelsels kunnen noemen, en hun oplossingen mogelijke oplossingen.
De vraag, die we ons nu stellen is:
als de vector y gegeven is, hoe is dan te zien of deze een mogelijke oplossing is. De volgende stelling doet hierover een uitspraak.

(24.14) Stelling. Er is een δA met $|\delta A| \leq \Delta A$ en er is een δb met $|\delta b| \leq \Delta b$ terwijl $(A + \delta A)y = b + \delta b$ d.e.s.d. als $|r| \leq \Delta A\,|y| + \Delta b$ met $r = Ay - b$.
bewijs :
$\Rightarrow$: $(A + \delta A)y - (b + \delta b) = r + \delta A y - \delta b = 0.$
coördinaatsgewijs: $r_i = -\Sigma\delta a_{i,j} y_j + \delta b_i \quad i = 1(1)n.$

$$\text{dus: } |r_i| \leq \Sigma|\delta a_{i,j}|\,|y_j| + |\delta b_i|$$

$$\text{of ook: } |r| \leq |\delta A|\,|y| + |\delta b|.$$

$$\text{dus: } |r| \leq \Delta A\,|y| + \Delta b.$$

$\Leftarrow$: $|r| \leq \Delta A\,|y| + \Delta b.$ Voor de i^{de} coördinaat r_i betekent dit:

$$-\sum_{j=1}^{n} \Delta a_{i,j}\,|y_j| - \Delta b_i \leq r_i \leq \sum_{j=1}^{n} \Delta a_{i,j}\,|y_j| + \Delta b_i.$$

Men gaat nu eenvoudig na dat er $\delta a_{i,j}$ en δb_i te kiezen zijn zo dat $r_i = -\sum_{j=1}^{n} \delta a_{i,j}\, y_j + \delta b_i \quad i=1(1)n$ en

$$|\delta a_{i,j}| < \Delta a_{i,j} \text{ en } |\delta b_i| < \Delta b_i.$$

(24.15) Voor de Höldernormen geldt:

1°. $\|x\|_p = \|\,|x|\,\|_p$
2°. als $0 < x < y$ dan $\|x\|_p < \|y\|_p$.
3°. $\|A\|_p \leq \|\,|A|\,\|_p$
4°. als $0 < A < B$ dan $\|A\|_p < \|B\|_p$.
Bewijs dit.

(24.16) Opmerking. Stelling (24.14) staat toe numeriek na te gaan of vooraf gekozen ΔA en Δb voldoen.
Zijn ΔA en Δb bekend, dan kan men vervolgens $\|\Delta A\|$ en $\|\Delta b\|$

bepalen in een geassocieerde Hölder norm en daarmee een bovengrens voor $\|\delta A\|$ en $\|\delta b\|$.

§ 24.3 Conditie getallen.

Wij gaan uit van (24.2): $\delta x = A^{-1}(\delta b - \delta A.x)$.
afschatten geeft:

$$\|\delta x\| \leq \|A^{-1}\|\{\ \|Ax\|\frac{\|\delta b\|}{\|b\|} + \|\delta A\|\ \|x\|\ \},$$

zodat voor de relatieve fouten geldt

(24.17) $$\frac{\|\delta x\|}{\|x\|} \leq \|A\|\ \|A^{-1}\|\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right).$$

Zij $\|A\|.\|A^{-1}\| = C(A)$, dan geeft $C(A)$ aan hoe de relatieve fout in A en b door kan werken in de relatieve fout in x.
C(A) noemt men het bij de gebruikte norm passende conditiegetal van A.

(24.18) Opgave. Toon aan $C(A) \geq 1$. Ga na dat $C(A) = C(\alpha A)$ voor α een scalair $\neq 0$.

(24.19) Opgave. Zij A unitair. Toon aan dat t.o.v. de euclidische norm geldt $C(A) = 1$.

(24.20) Opgave. Zij A niet singulier. Toon aan dat t.o.v. de euclidische norm geldt $C(A) = \frac{\max\sqrt{|\lambda(A^*A)|}}{\min\sqrt{|\lambda(A^*A)|}}$. aanw: $\|A^{-1}\| = \frac{1}{\inf_{\|x\|=1} \|A x\|}$.

(24.21) We gaan de meetkundige betekenis na van het conditiegetal passend bij de 2-norm (= euclidische norm). Zij A_i steeds de i-de kolom van de matrix A, A_i opgevat als vector uit de $\mathbb{R}^n$. Evenzo zij $\tilde{A}_i$ de gespiegelde van de i^{de} rij van A^{-1} opgevat als vector van de $\mathbb{R}^n$.
Dan geldt:

(24.22) Stelling. Zij a_i de euclidische afstand van A_i tot het hypervlak opgespannen door de overige A_j en zij A niet singulier
Dan is $a_i = \frac{1}{\|\tilde{A}_i\|_2}$.

Bewijs: We nemen eenvoudigheidshalve $i = 1$.
De afstandsvector staat loodrecht op $A_2, \dots, A_n$. Wegens $\tilde{A}_1^T A_j = 0$ voor $j \neq 1$ is de afstandsvector dus een veelvoud van $\tilde{A}_1$.

Stel $A_1 = \lambda \tilde{A}_1 + \sum_{j=2}^{n} \lambda_j A_j$.

dan geldt: $a_1 = |\lambda| . \|\tilde{A}_1\|_2$.

Uit $\tilde{A}_1^T A_1 = 1$ volgt $|\lambda| \, \|\tilde{A}_1\|_2^2 = 1$.

Dus $|\lambda| = \dfrac{1}{\|\tilde{A}_1\|_2^2}$ en $a_1 = \dfrac{1}{\|\tilde{A}_1\|_2}$.

(24.23) Stelling. Voor alle i en een zekere j geldt:

$$\left(\frac{a_i}{\|A\|_2}\right)^{-1} \leq C(A) \leq \sqrt{n}\left(\frac{a_j}{\|A\|_2}\right)^{-1}$$

$\tilde{A} = A^{-1}$

Bewijs $C(A) = \|A\|_2 . \|\tilde{A}\|_2$.

$$\left(\frac{a_i}{\|A\|_2}\right)^{-1} = \|A\|_2 . \|\tilde{A}_i\|_2 .$$

Nu geldt voor alle i $\|\tilde{A}_i\|_2 \leq \|\tilde{A}\|_2$ en is er een j zodat

$$\|\tilde{A}_j\|_2 \geq \frac{1}{\sqrt{n}} \|\tilde{A}\|_F \geq \frac{1}{\sqrt{n}} \|\tilde{A}\|_2 .$$

(24.24) Als A_i een hoek θ_i maakt met het hypervlak door de overige A_j geldt $C(A) \geq \dfrac{1}{\sin(\theta_i)}$. Ga na. De hoek θ_i is ook de hoek tussen de beelden onder de matrix A van twee orthogonale vectoren, nl. de i^{de} basis-vector en een zekere lineaire combinatie van de overige basisvectoren.
Zij nu θ de minimale hoek tussen de beelden van twee orthogonale vectoren, het minimum genomen over alle paren orthogonale vectoren.
Dan

(24.25) $\theta = 2\times \text{arc cotg} (C(A))$
Dit volgt uit de Wielandt-ongelijkheid, geformuleerd en bewezen in F.L. Bauer, A.S. Householder: Some inequalities involving the euclidean condition of a matrix. Num. Math 2. 1960.
Een meer elementair bewijs is verwerkt in de volgende opgaven.

(24.26) Opgave. Zij A symmetrisch en niet-singulier
a. Toon aan dat θ = hoek (Ax_0, Ax_1) met

$x_0 = e_{max} - e_{min}$ en $x_1 = e_{max} + e_{min}$, waarbij e_{max}, e_{min} eigenvectoren van A, e_{max} bij λ_{max}, e_{min} bij λ_{min}.

$|\lambda_{max}| = \max|\lambda(A)|$,

$|\lambda_{min}| = \min|\lambda(A)|$. $\|e_{max}\| = \|e_{min}\|$

b. bewijs voor dit geval (24.25).

(24.27) Opgave. Zij A non-singulier. x_0 en x_1 zijn als in (24.26) gedefiniëerd bij de matrix A^*A.

a) Toon aan θ = hoek (Ax_0, Ax_1), θ behorend bij A.

b) Bewijs (24.25).

(24.28) We geven nog enige gevolgen van (24.23).
Noemen we een stelsel vectoren bijna-afhankelijk als één van hen een kleine hoek maakt met het vlak door de overige dan impliceert bijna-afhankelijkheid der A_i dus een groot conditiegetal.

(24.29) Een groot conditiegetal impliceert echter niet bijna-afhankelijkheid (vgl.(24.28)), zoals men inziet aan de hand van een orthogonale matrix waarin één kolom met 10^8 is vermenigvuldigd.

(24.30) Een groot conditiegetal impliceert echter wel bijna-afhankelijkheid als de kolommen van A ongeveer gelijke norm hebben. Laat nl. de kolommen van A maar precies gelijke norm hebben, dan is $\|A\|_2 \leq \|A\|_F = \sqrt{n}\|A_j\|$ voor elke j zodat uit (24.23) volgt dat voor zekere j geldt $C(A) \leq \frac{n}{\sin(\theta_j)}$.

(24.31) Opmerking. Bij andere normen zal de interpretatie van het conditiegetal natuurlijk anders zijn. Men bedenke echter dat de verhouding van verschillende normen op een eindig dimensionale ruimte begrensd is. Met name voor de 1-, 2- en ∞-normen en voor de gebruikelijke waarden van n (zeg $n < 100$) zijn deze grenzen $<$ 10 of 100 (zie (20.8)). Waar het hier slechts om kwalitatieve beschouwingen gaat zal overgang op andere normen dus geen essentiële wijzigingen veroorzaken, en zullen factoren als de genoemde ons weinig belang inboezemen.

(24.32) Gewoonlijk zullen schattingen voor $\|\delta x\|$, m.b.v. (24.17) uitgaan-

de van δA en δb ernstige overschattingen **zijn.** Zo geldt b.v. voor $\delta A = \lambda A$: $\frac{\|\delta x\|}{\|x\|} = \frac{\|\delta A\|}{\|A\|}$, ongeacht het conditiegetal van A. (24.17) is echter wel scherp in de zin van stelling (24.9)/ Het conditiegetal is slechts dan een goede maat voor de mogelijke versterking van de relatieve fout als alle δA van gegeven norm mogelijk zijn.

(24.33) Opgave

$$A = \begin{bmatrix} 100 & 0 & 10 \\ 0 & 100 & -40 \\ 10 & -40 & 18 \end{bmatrix} \qquad b = \begin{bmatrix} 1,1 \\ 0 \\ 10 \end{bmatrix}.$$

In twee decimalen nauwkeurig rekenend vindt men als opl. van $Ax = b$ de vector x^*.

$$x^{*T} = (-0,88\ ,\ 3,6,\ \ 8,9).$$

Geef een bovengrens voor $\|x - x^*\|$.

Ga na, dat men mag aannemen dat $A + \delta A$ weer **symmetrisch** is en nullen heeft waar A nullen heeft.

/ t.a.v. δA, niet t.a.v. δb

§ 25 Numerieke uitvoering van het proces van Gauss - eliminatie.

(25.1) In de eerste stap Gauss - eliminatie wordt bij de 2^e t/m n^e rij een vector in de richting van de eerste rij opgeteld. Als deze vector "groot" is t.o.v. de rij in kwestie (b.v. als $a_{11}^{(1)}$ klein t.o.v. $a_{i1}^{(1)}$ terwijl de overige elementen van de 1^e en i^e rij dezelfde grootte orde hebben, (vgl.(21.8)) dan heeft deze rij na bijtelling bijna de richting van de eerste rij m.a.w. in deze situatie wordt het stelsel bijna - afhankelijk gemaakt. Aangezien men kan verwachten dat bijna - afhankelijke stelsels gevoeliger zijn voor perturbaties (en daarmee wellicht ook voor afrondfouten) dan niet bijna - afhankelijke stelsels is dit ongewenst. (vgl. (24.28)).

(25.2) Voorbeeld van bijtelling "grote vector":

$$0{,}00067\, x_1 + x_2 = 1{,}999$$
$$0{,}5\, x_1 + 2x_2 = 3$$

Rekent men in 4 decimalen (vgl. § 4) dan luidt het stelsel na 1 stap Gauss - eliminatie met a_{11} als pivot:

$$0{,}00067\, x_1 + x_2 = 1{,}999$$
$$-744{,}3x_2 = -1489$$

zodat $x_1 = -2{,}985$

$x_2 = 2{,}001$

De exacte uitkomsten (afgerond op 4 decimalen) zijn echter

$x_1 = -2{,}001$

$x_2 = 2{,}000$

(25.3) Opgave. Ga na in (25.2) dat als $x_2 = 2 + \varepsilon$, uit de eerste vgl. volgt: $x_1 = -1{,}493 - 1493.\varepsilon$.

(25.4) Voorbeeld van een stelsel dat zelf al bijna afhankelijk is: Passen we als in (25.2) Gauss - eliminatie toe op het stelsel

$$1{,}01\, x_1 + 3{,}4\, x_2 = 7{,}82$$
$$0{,}34\, x_1 + 1{,}1\, x_2 = 2{,}61$$

dan vindt men, rekenend in 4 decimalen, $x_1 = 3{,}432$, $x_2 = 1{,}280$.
De exacte antwoorden (afgerond op 4 decimalen): $x_1 = 3{,}326$, $x_2 = 1{,}312$.

(25.5) Het ligt dus voor de hand als eerste pivot niet zo maar een $a_{i1}^{(1)}$ te kiezen, die $\neq 0$ is (zie (21.6)), maar er naar te streven dit zo te doen dat de t.g.v. het eliminatieproces bij een rij opgetelde vector niet groot is t.o.v. deze rij.

Dit kan als volgt gerealiseerd worden.
Vermenigvuldig alle vergelijkingen met zodanige factoren dat in de coëfficiëntenmatrix elke rij ongeveer gelijke norm θ heeft (in een willekeurige, maar vast gekozen norm).
Neem nu als pivot $a_{i1}^{(1)}$ met $|a_{i1}^{(1)}| = \max_k |a_{k1}^{(1)}|$ (zie 21.6) en verwissel de eerste en de i^e vergelijking.

(25.6) Ga na dat bij de eliminatie van x_1 aldus bij elke rij een vector wordt opgeteld van hoogstens norm θ.

(25.7) Het brengen van de matrixrijen op ongeveer gelijke norm noemt men <u>rij - equilibratie</u>.

(25.8) <u>Opmerking</u>. Wanneer men in een ongeëquilibreerde matrix het absoluut grootste kolom element als pivot neemt, behoeft dit niet het beoogde effect te hebben. (Ga na).

(25.9) Bij de eliminatie van $x_2, \ldots$ volstaat men met pivoting (dus hier het zoeken van de absoluut grootste $a_{ik}^{(k)}$, $i > k$, plus de bijbehorende verwisseling van rijen) en laat men equilibratie achterwege.

(25.10) Dat het eliminatie proces ook zonder tussentijdse equilibratie "stabiel" is (in die zin dat bij de rijen van de oorspronkelijke matrix geen willekeurige grote combinaties van de overige rijen worden opgeteld) kan men als volgt inzien.

(25.11) Stel dat de vergelijkingen reeds zo gerangschikt staan dat tijdens het proces <u>zonder</u> tussentijdse rijequilibratie
$\max_{i>k} |a_{ik}^{(k)}| = |a_{kk}^{(k)}|$ voor alle k, d.w.z. dat de pivot steeds op de juiste plaats verschijnt en geen rijverwisseling nodig is.
Zij $r_i^{(k)}$ de i^e rij van $A(k)$.

(25.12) $r_i^{(k)}$ ontstaat uit $r_i^{(1)}$ door bijtelling van een vector
$$\sum_{j=1}^{k-1} c_{ji} r_j^{(j)} \quad , \quad |c_{ji}| \leq 1. \ (i > k)$$

(25.13) Met inductie ziet men in dat de bijgetelde vector een norm $\leq (2^{k-1}-1)\,\theta$ heeft (vgl. ook (26. 11)), hetgeen voor grotere waarden van k erg kan oplopen. (Rekenen we b.v. in 13 decimalen dan kan blijkbaar voor $k \gtrsim 40$ $r_i^{(1)}$ geheel ondergaan in wat er bijgeteld wordt). In elk geval is wat er bij een rij opgeteld wordt begrensd. En in werklijkheid zal het natuurlijk zeer zeldzaam zijn dat inderdaad $\|\sum_{j=1}^{k-1} c_{ji} r_j^{(j)}\| = (2^{k-1}-1)\theta$. Maar het kan wel:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 1 \\ -1 & -1 & 1 & 0 & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{pmatrix}$$

(25.14) Overigens is een goede vergelijking tussen het proces met en zonder tussentijdse equilibratie moeilijk, en nog niet uitgevoerd.

(25.15) Het hierboven beschreven proces, waarbij men alleen aan het begin equilibreert en pivot op de aangeduide manier is wel het meest gebruikte proces om vergelijkingen op te lossen. Men noemt het: Gauss - eliminatie met equilibratie en partial pivoting (partial in tegenstelling tot complete, welk geval men niet het absoluut grootste element in de aan de beurt zijnde kolom opzoekt maar het absoluut grootste van de hele nog te bewerken rechtsondermatrix).

§ 26 Het effekt van afrondfouten

(26.1) Het door berekening verkregen resultaat van een proces is vaak te interpreteren als het exacte resultaat behorend bij enigzins verstoorde beginwaarden (vgl. §24.1). Dit is ook het geval bij het Gauss - eliminatieproces met partial pivoting. Een uitvoerige analyse is gemaakt door Wilkinson die voor dit proces, evenwel ongeacht of de matrix geëquilibreerd is, heeft aangetoond:

(26.2) Stelling. De via Gauss - eliminatie met **partial** pivoting berekende oplossing van het stelsel $Ax = b$ voldoet exact aan:

$$(A + \Delta A)y = b$$

met

$$\frac{\|\Delta A\|_\infty}{\|A\|_m} \leq 1{,}01 .\ \phi(n).\ g(A).\xi$$

Waarin $\phi(n) \leq n^3 + 3n^2$

$\|A\|_m = \max |a_{ij}|$

$$g(a) = \frac{\max\limits_{i,j,k,} \left|a_{i,j}^{(k)}\right|}{\|A\|_m}$$

(26.3) De grootheid g(A) geeft dus aan hoeveel maal zo groot het grootste element te eniger tijd in het eliminatieproces is als het grootste element van de oorspronkelijke matrix; g(A) is dus een soort groeifactor. Merk op dat $g(A) \geq 1$ wegens $a_{ij} = a_{ij}^{(1)}$

(26.4) Uit (24.4) volgt nu:

(26.5) $\frac{\|\delta x\|_\infty}{\|x\|_\infty} \leq \|A^{-1}\|_\infty . \|\Delta A\|_\infty \leq \|A^{-1}\|_\infty \|A\|_m . 1{,}01 . \phi(n) . g(A) . \xi$

(26.6) De grootheid $c'(A) = \|A^{-1}\|_\infty . \|A\|_m$ is weer een soort conditie getal.

(26.7) <u>Opgave</u>. Ga na $C'(A) \geq 1$, $C'(\lambda A) = C'(A)$, en

$$C'(AB) \geq \frac{1}{n} \max\left\{\frac{C'(A)}{C'(B^{-1})}, \frac{C'(B)}{C'(A^{-1})}\right\}$$

(Hint: bedenk dat $n\|A\|_m$ een multiplicatieve matrix norm is.)

(26.8) We bezien de invloed van rijequilibratie op C'(A). Ga na dat rijequilibratie van een matrix neerkomt op linksvermenigvuldiging met een geschikte diagonaalmatrix.

(26.9) <u>Stelling</u>. C'(DA) is minimaal als in DA alle rijen gelijke ∞ - normen hebben.

<u>Bewijs</u>.

Stel dat in A reeds alle rijen gelijke ∞ - norm θ hebben. Dan is $\|DA\|_m = \|D\|_m . \theta = \|D\|_m . \|A\|_m$

Wegens $\|A^{-1}D^{-1}D\|_\infty \leq \|A^{-1}D^{-1}\|_\infty . \|D\|_\infty$ geldt:

$C'(DA) = \|A^{-1}D^{-1}\|_\infty . \|DA\|_m \geq \|A^{-1}\|_\infty . \|A\|_m . \frac{\|D\|_m}{\|D\|_\infty} = C'(A).$

(26.10) Derhalve: equilibratie minimaliseert de factor $C'(A)$ in het rechterlid van (26.5). Nemen we aan dat A al geëquilibreerd is dan kunnen we zeggen dat geen enkele andere rijschaling een rechterlid voor (26.5) kan opleveren dat meer dan een factor $g(A)$ kleiner is dan voor A zelf (immers $g \geq 1$ voor elke matrix).

(26.11) De vraag is nu nog wat equilibratie met $g(A)$ doet. We merken op dat t.g.v. partial pivoting $g(A) \leq 2^{n-1}$. Om dit in te zien beschouwen we het elimineren van x_1. Stel dat $|a_{11}| = \max_k |a_{k1}|$ (zie (25.5)) zodat men a_{11} als pivot gebruikt. Dan wordt bij elke volgende rij hoogstens de 1e rij opgeteld of afgetrokken, zodat $\max_{i,j} |a_{i,j}^{(2)}| \leq 2 \max_{i,j} |a_{ij}|$.

Analoog bij de volgende eliminatie stappen. Men kan zelfs een matrix aangeven waarbij 2^{n-1} wordt gehaald. Men zal echter begrijpen dat dit een uiterst ongelukkige samenloop van omstandigheden vereist, en gelukkig blijkt in de praktijk slechts hoogst-zelden $g(A) > 4$, <u>of de matrix geëquilibreerd is of niet</u> (het "wonder van Wilkinson"; Wilkinson heeft dit ontdekt)

(26.12) Men kan derhalve zeggen dat equilibratie vrijwel steeds het rechterlid van (26.5) tot zijn minimum reduceert op hoogstens een factor 4 na. Dit is een indicatie voor het praktische succes van de methode.

(26.13) Merk op dat men tijdens de berekening kan vaststellen wat $g(A)$ is. Mocht men $g(A)$ te groot vinden, dan kan men overstappen op andere oplosmethoden. Deze, alsmede vele nadere details van het voorafgaande vallen echter buiten het bestek van dit college.

Filosofie bij (26.11): Gauss met pivoting is 2x zo goedkoop als stabielere processen. Gauss is net zo goed mits de groeifactor klein blijft. Als de groeifactor niet klein blijft stap je over op een stabielere methode. Dat hoeft maar hoogst zelden. Dus zo is de economie gediend

§ 27 Naverfijning

(27.1) Met x_i e.d. (i=1,2,...) zullen we hier het i^e-element van een rij van vectoren aanduiden en niet een coördinaat.

(27.2) Stel dat we via Gauss - eliminatie met partial pivoting voor $Ax = b$ een benaderde oplossing x_1 vinden.
Enige verbetering van deze waarde is soms mogelijk door nauwkeurige berekening van het residu $r_1 = b - Ax_1$ en als nieuwe benadering te nemen:

$$x_2 = x_1 + v_1$$

waarin v_1 de berekende oplossing is van $Az = r_1$. Bedenk dat $z = v_1$ nu snel bepaald kan worden omdat de LU - decomp. van het stelsel na eenmaal oplossen van $Ax = b$ bekend is (zie ook (22.15))
Het proces kan men herhalen met x_2 etc.

(27.3) Er wordt zo een rij $x_1, x_2, x_3, \ldots, x_k, x_{k+1}, \ldots$ van benaderde oplossingen geconstrueerd waarbij $x_{k+1} = x_k + v_k$ en v_k de met Gauss - eliminatie berekende oplossings vector is van het stelsel $Az = r_k = b - Ax_k$.

(27.4) De vraag is nu: convergeert de rij x_i naar de exacte oplossing van het oorspronkelijke stelsel?

(27.5) Wegens (26.2) kunnen we aannemen dat v_k exact voldoet aan een vergelijking:

$$(A + (\Delta A)_k)\, v_k = r_k$$

Stellen we $B_k = A^{-1}(\Delta A)_k$ dan geldt:

(27.6)
$$A(I + B_k)\, v_k = r_k$$

(27.7) Stelling. Laat in (27.6) voor alle k $B_k < \varepsilon < \frac{1}{2}$ in zekere vastgekozen geassocieerde norm. Dan convergeert de rij x_k naar de exacte oplossing van het stelsel $Ax = b$.
Bewijs Zij α exacte wortel van $Ax = b$, d.w.z. $\alpha = A^{-1}b$. Laat

$h_k = \alpha - x_k$

$x_{k+1} = x_k + v_k$

$\Rightarrow A(I + B_k)\, x_{k+1} = A(I + B_k)\, x_k + r_k = AB_k\, x_k + b$

$\Rightarrow (I + B_k)\, x_{k+1} = B_k x_k + \alpha$

$\Rightarrow (I + B_k)\, h_{k+1} = B_k\, h_k$

Wegens $\|B_k\| < 1$ is $(I + B_k)$ nonsingulier((19.24)) zodat:

$$h_{k+1} = (I + B_k)^{-1}. B_k h_k$$

$$\Rightarrow \|h_{k+1}\| \leq \|(I + B_k)^{-1}\|. \|B_k\|. \|h_k\|$$

Met (20.10) volgt:

$$\|h_{k+1}\| < \frac{\|B_k\|}{1-\|B_k\|} \|h_k\|$$

zodat voor $\|B_k\| \leq \varepsilon < \frac{1}{2}$ geldt:

$$\|h_{k+1}\| \leq \frac{\varepsilon}{1-\varepsilon} \|h_k\| < \|h_k\|.$$

(27.8) Corollarium. Naverfijning bij Gauss - eliminatie met partial pivoting convergeert als

$$1{,}01.\ C'(A).\ \phi(n).\ g(A).\ \xi < \tfrac{1}{2}$$

Bewijs. (26.2), (27.5), (27.6) en (27.7).

(27.9) Opmerkingen.

(27.10) Er is aangenomen dat arithmetische fouten alleen optreden tijdens het oplosproces en r_k en $x_k + v_k$ 'exact' berekend worden.
Rekenen in enkelvoudige precisie is dan niet toereikend. De afrondfouten bij de berekening van inprodukten zullen de evaluatiefout in $r_k = b - Ax_k$ (relatief) aanzienlijk doen oplopen. Men mag niet meer hopen dat de oplossing van dit stelsel $Az = r_k$ (met berekende r_k) een goede correctie van x_k oplevert.
Gebruikt men evenwel een dubbellengte - procedure voor berekening van inprodukten, dan krijgt men r_k voldoende nauwkeurig (althans voorlopig).
Het berekende residu dient als rechterlid in de vergelijking $Ax = r_k$ waarvan de oplossing v_k behept zal zijn met een rel.fout $\sim \kappa.\ C'(A)\ \xi$ (κ zekere constante). (vgl. (26.5)).
Mits aan de convergentie conditie van het proces is voldaan zullen de benaderde oplossingen x_k aldus aanvankelijk op een adequate wijze door de v_k worden gekorrigeerd.
In het gunstige geval treedt convergentie op d.w.z. zullen vanaf zeker moment de decimalen van x_k niet meer gewijzigd worden. Grotere nauwkeurigheid is dan alleen mogelijk als ook de addities $(x_k + v_k)$ in dubbele precisie worden uitgevoerd.
Het kan echter ook gebeuren dat na aanvankelijke verbetering van

er is geen pag 93

de benaderde oplossingen verdere correcties een oscillatie der x_k doen ontstaan. Het is niet te verwachten dat optellen in dubbele lengte dan nog noemenswaardige verbetering brengt. Een foutenanalyse voor het naverfijningsproces vindt men b.v. in C.B. Moler-JACM 14 (1967) 316 - 321

(27.11) Uit het bewijs van (27.7) volgt nog een indruk van de converg ntiesnelheid. Per iteratiestap vermindert de fout met een factor $\lesssim \frac{\varepsilon}{1-\varepsilon}$.

Ook hier is men kennelijk gebaat bij equilibratie van de matrix, omdat dan C'(A) en daarmee dus in zekere zin de convergentiefactor bij naverfijning (zie (27.8)) geminimaliseerd wordt.

§ 28 Iteratieve methoden.

(28.1) Iteratieve methoden (vgl. §11) ter oplossing van $Ax = b$ berusten veelal op een geschikte splitsing $A = A_1 + A_2$ waarna men het stelsel $x = \phi(x) = A_1^{-1}b - A_1^{-1}A_2x$ door successieve substitutie tracht op te lossen. Het is daarbij van belang dat A_1^{-1} zonder al te veel moeite bepaald kan worden.

(28.2) Voorbeelden.

(28.3) Gauss - Jacobi iteratie: neem A_1 een diagonaalmatrix met dezelfde diagonaal als A.

(28.4) Gauss - Seidel iteratie: neem A_1 een benedendriehoeksmatrix met dezelfde benedendriehoek als A.

(28.5) Opgave. Toon aan dat de rij $x_1, x_2, \ldots$ met $x_{i+1} = A_1^{-1}b - A_1^{-1}A_2x_i$ convergeert naar de wortel van $Ax = b$ indien in zekere geassocieerde norm $\|A_1^{-1}A_2\| < 1$.

(28.6) Men kan aantonen dat Gauss - Jacobi en Gauss - Seidel convergeren voor diagonaaldominante matrices, Gauss - Seidel ook voor symmetrische positief - definiete matrices.
Een bezwaar van de iteratieve methoden is dat zij doorgaans slechts langzaam convergeren.
Later in dit college zullen we nog wel nadere bijzonderheden over deze processen ontmoeten. Men zie ook b.v. Varga (ch. 3).

§ 29 Lineaire kleinste kwadraten; Inleiding

Een belangrijk probleem in de lineaire algebra is het "zo goed mogelijk oplossen" van een overbepaald stelsel $Ax = b$, d.i. een stelsel waarin A een $m \times n$ matrix is, $m > n$ (een dergelijk stelsel zal immers i.h.a. geen echte oplossing hebben) Dit probleem ontmoet men bijv. bij het uitwerken van de resultaten van experimenteel onderzoek, waar men voor een aantal onbekende grootheden $x_1, \ldots x_n$ en groot aantal lineaire relaties $\sum_j a_{ij} x_j = b_i$, $i = 1,2,\ldots,m$ kan bepalen, waarvan de coëfficiënten echter niet erg nauwkeurig zijn. Of men wenst een functie, die men in m punten kent, te vervangen door een (hoogstens) $(n-1)$-ste graads polynoom.
Onder de "zo goed mogelijke oplossing" verstaan we nu die x waarvoor $\|Ax - b\|$ minimaal is in een of andere norm. Met de 2-norm noemt men dit een kleinste kwadraten oplossing, omdat voor deze x de uitdrukking $\sum_i [\sum_j a_{ij} x_j - b_i]^2$ geminimaliseerd wordt, met de ∞ - norm een Chebychev - oplossing. In de volgende paragrafen beperken we ons uitsluitend tot kleinste kwadraten oplossingen. Bovendien veronderstellen we dat de lineaire ruimten reëel zijn.

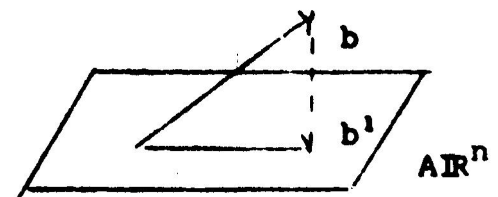
§ 30 Normaalvergelijkingen (van Gauss).

(30.1) De klassieke aanpak van dit probleem is met de zgn. normaalvergelijkingen van Gauss. Dat gaat als volgt:

(30.2) Zij A een $m \times n$ matrix, $m > n$, met rang $(A) = n$.

(30.3) Definitie. $x \in \mathbb{R}^n$ heet een kleinste kwadraten oplossing van $Ax = b$ indien voor alle $y \in \mathbb{R}^n$: $\|Ax - b\|_2 \leq \|Ay - b\|_2$.

(30.4) Nu beeldt A de $\mathbb{R}^n$ af op een n-dimensionale lineaire deelruimte van de $\mathbb{R}^m$, die wordt opgespannen door de kolommen (als vectoren beschouwd) van A. Blijkbaar moet Ax dan die vector in $A\mathbb{R}^n$ zijn, die de kleinste 2-afstand tot b heeft, d.i. de projectie b' van b op $A\mathbb{R}^n$.
Dus $(A\mathbb{R}^n, Ax - b) = 0$,
of $(\mathbb{R}^n, A^*Ax - A^*b) = 0$.
We hebben nu bewezen:



(30.5) <u>Stelling</u>: x is een kleinste kwadraten oplossing van $Ax = b$ als $A^*A\,x = A^*b$.

(30.6) <u>Opgave</u>. Bewijs ook rechtstreeks dat $\|Ay - b\|_2 > \|Ax - b\|_2$ als $A^*Ax = A^*b$ en $y \neq x$.

(30.7) Het lineaire stelsel vergelijkingen $A^*Ax = A^*b$ noemt men de normaal vergelijkingen van Gauss.

(30.8) <u>Opgave</u>. Als A rang n heeft is A^*A positief definiet en dus niet singulier. In dat geval kan men dus spreken van <u>de</u> kleinste kwadraten oplossing. Als A rang $< n$ heeft is de kleinste kwadraten oplossing niet meer eenduidig bepaald; echter is Ax voor alle kleinste kwadraten oplossingen hetzelfde. Ga dit alles na.

(30.9) Het lineaire stelsel $A^*Ax = A^*b$ kan men in het niet singuliere geval wegens (30.8) oplossen m.b.v. Choleski.

(30.10) De aanpak met Normaalvergelijkingen werkt bevredigend voor niet te grote waarden van n (rg $n \leq 5$), hetgeen in de tijd van het rekenen met de hand steeds het geval was. Met de komst van de computer werden veel grotere waarden van n mogelijk en toen bleken zeer onaangename verschijnselen. We laten dit eerst zien voor het (veelvoorkomende) geval van polynoomaanpassing, en tonen hoe het wél moet. Later zullen we dit ook voor het algemenere probleem doen.

§ 31 <u>Polynoomaanpassingen</u>.

(31.1) Op $[0,1]$ is een functie getabuleerd gegeven in de punten $t_1, \ldots t_m$, homogeen verspreid op het segment $[0,1]$. Gevraagd wordt een $(n-1)$ste graads kleinste kwadraten polynoom aanpassing, $n-1 < m$.

(31.2) Schrijf het gevraagde polynoom als:

(31.3) $a_0 + a_1x + \ldots + a_{n-1}x^{n-1}$, waarin $a_0, \ldots, a_{n-1}$ de onbekenden zijn.

(31.4) Met $A = \begin{pmatrix} 1 & t_1 \cdots\cdots & t_1^{n-1} \\ \vdots & \vdots & \vdots \\ 1 & t_m \cdots\cdots & t_m^{n-1} \end{pmatrix}$ en $x = \begin{pmatrix} a_0 \\ \vdots \\ a_{n-1} \end{pmatrix}$ hebben

we dus een lineair kleinste kwadraten probleem $Ax = b$, waarin b de m- dimensionale vector van functiewaarden is.

(31.5) De matrix coëfficiënten van de matrix A^*A worden dan gegeven door

(31.6) $$(A^*A)_{i,j} = \sum_{k=1}^{m} (t_k)^{i+j-2}. \quad \text{(Ga na)}$$

(31.7) Voor m groot zal gelden: $(A^*A)_{i,j} \approx m \int_0^1 t^{i+j-2} dt = \frac{m}{i+j-1}$. Ga na

(31.8) Gevolg: voor grote m zal de matrix A^*A sterk gelijken op de matrix

$$m \times \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \cdots\cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} \cdots\cdots\cdots & & \frac{1}{n+1} \\ \vdots & & & \\ \frac{1}{n} & \frac{1}{n+1} \cdots\cdots\cdots & & \frac{1}{2n-1} \end{pmatrix}.$$

Zonder de factor m is dit een eindig segment van de Hilbert-matrix. (De Hilbert-matrix is de oneindig voortlopende matrix).
Men kan aantonen, dat het conditiegetal van het n x n - segment $\sim 2^{5n}$ is. Binair rekenend met 40 - bits kan alleen al het afronden van de coëfficiënten van A^*A vanaf $n = 9$ de oplossing onherkenbaar verminken.

(31.9) De juiste oplossingswijze voor dit soort kleinste kwadraten problemen loopt over orthogonale polynomia. We geven hiervan een korte schets. Voor details zie TA §8 e.v.

§ 32 Orthogonale polynomia.

(32.1) Bij het oplossen van het in §31 gestelde probleem hebben we geeist, dat de oplossing de gedaante

$$a_0 + \ldots + a_{n-2}\, x^{n-2} + a_{n-1}\, x^{n-1} \quad \text{had.}$$

We kunnen, algemener, eisen dat de oplossing de gedaante

(32.2)
$$a_0\, p_0(x) + \ldots + a_{n-1}p_{n-1}(x)$$

heeft met p_i een polynoom van precies graad i. Geheel analoog aan (31.4) kan men weer de formulering als kleinste kwadraten probleem geven. Ga na dat nu $A = \begin{pmatrix} p_0(t_1) \ldots\ldots p_{n-1}(t_1) \\ \ldots\ldots\ldots\ldots\ldots \\ p_0(t_m) \ldots\ldots p_{n-1}(t_m) \end{pmatrix}$, waarbij $t_1,\ldots t_m$ de gegeven steunpunten zijn.

We zullen trachten de polynomen p_i zo te bepalen dat A^*A een prettige gedaante heeft.

(32.3) De matrix coëfficiënten van A^*A worden

(32.4)
$$(A^*A)_{i,j} = \sum_{k=1}^{m} p_i(t_k)\, p_j\,(t_k).$$

(32.5) Een matrix die beslist niet aan het in §31 gesignaleerde euvel mank gaat is vanzelfsprekend de eenheidsmatrix. Dat betekent dus dat de polynomia $\{p_i\}$ voldoen aan:

(32.6)
$$\sum_k p_i(t_k)p_j(t_k) = \delta_{ij}$$

(32.7) Nu vormen de polynoom van graad $\leq m-1$ een lineaire ruimte P_{m-1}. (Ga na!).

(32.8) <u>Opgave</u>. Zij voor $i = 0,\ldots,m-1$ p_i een polynoom van precies graad i. Toon aan dat de $\{p_i\}$ een basis vormen van P_{m-1}.

(32.9) Voor m gegeven punten $t_1,\ldots,t_m$ en q_1, $q_2 \in P_{m-1}$ definiëert

(32.10)
$$(q_1, q_2) = \sum_{k=1}^{m} q_1\,(t_k)\, q_2(t_k)$$

een inprodukt in de ruimte P_{m-1}.

(32.11) <u>Opgave</u>. Toon dit aan.

(32.12) Blijkbaar impliceert nu onze wens $A^*A = I$ dat

$$\left.\begin{array}{ll} (p_i,p_i) = 1 & \\ (p_i,p_j) = 0 & i \neq j \end{array}\right\} \; 0 \leq i,j \leq m-1.$$

(32.13) Indien (32.12) geldt en p_i precies de graad i heeft zegt men dat de polynomen p_i een <u>orthogonaal stelsel</u> vormen t.o.v. het gebruikte inprodukt.

(32.14) Uit de lineaire algebra is bekend hoe men in een lineaire ruimte met inprodukt uit een gegeven basis een orthogonale basis construeert. (het proces van Gramm - Schmidt). Voor de polynomen p_i ligt de situatie eenvoudiger, zoals men ziet in TA §8 e.v.

(32.15) We kunnen de oplossing van ons probleem (het (n-1) ste graads polynoom vinden dat een in m punten gegeven functie in de zin van **kleinste kwadraten** benadert) nu expliciet aangegeven.
We beschouwen daartoe de m gegeven functiewaarden in $t_1, \ldots, t_m$ afkomstig van een polynoom f uit P_{m-1} (waarom kan dit?).
Dan geldt voor de kl. kw. oplossing x:

(32.16) $$x = ((p_0, f), (p_1 f), \ldots, (p_{n-1} f)).$$

en het bijbehorende kl. kw. polynoom is

(32.17) $$(p_0, f)\, p_0 + (p_1, f) p_1 + \ldots + (p_{n-1}, f)\, p_{n-1}.$$

(32.18) Opgave. Voor n = m is de kl.kw. opl. met het Lagrange - interpolatie polynoom.

(32.19) Wie dit practisch wil toepassen wordt aangeraden TA §8 e.v. te raadplegen, waar ook uiteengezet wordt hoe men deze zaken handig programmeert (met name het gebruikmaken van een drietermsrecursie blijkt zeer profijtelijk). Overigens zijn er standaard-procedures, die volgens deze lijnen werken, beschikbaar.

§33 De methode van Householder.

(33.1) In deze paragraaf zullen we een methode bekijken ter verkrijging van een kleinste kwadraten oplossing van het overbepaalde lineaire stelsel $Ax = b$ die stabieler is dan de methode van de normaalvergelijkingen.
Het proces berust op het vinden van een orthogonale matrix Q zodat $QA = U$, met U een bovendriehoeksmatrix.

(33.2) Allereerst tonen we aan dat zo'n decompositie bestaat. We geven een constructief bewijs waarbij de kolommen van de matrix Q achtereenvolgens worden berekend.
Centraal in dit bewijs staat de Householder - transformatie.

(33.3) Definitie. Voor elke $v \in \mathbb{R}^m$, $v \neq 0$, definiëren we de lineaire afbeelding

$$H_v = I - 2\frac{vv^T}{\|v\|_2^2}$$

We noemen deze afbeelding de bij v behorende Householder transformatie.

(33.4) Bewijs nu zelf de eigenschappen (33.5) t/m (33.8)

(33.5) $H_{\alpha v} = H_v$ voor elke constante α.

(33.6) Een Householder transformatie is orthogonaal en symmetrisch.

(33.7) Voor $u \perp v \Rightarrow H_v u = u$.

(33.8) Voor $u \perp v \Rightarrow H_v(u+v) = u - v$. Dus H_v is een spiegeling aan vlak door $0 \perp v$.

(33.9) Bij elke a en $b \in \mathbb{R}^m$, $a \neq 0$, $b \neq 0$ is er een $v \in \mathbb{R}^m$ zodat $H_v a$ de richting heeft van b, d.w.z. er is een constante λ zodat $H_v a = \lambda b$ en wel voldoet $v = a - \lambda b$ met $\lambda = \pm \|a\|_2 / \|b\|_2$.
Bewijs: Wegens (33.6) zal moeten gelden $\lambda = \pm \|a\|_2/\|b\|_2$ omdat, vanwege de orthogonaliteit $\|H_v a\|_2 = \|a\|_2$. Wegens (33.8) is het voldoende orthogonale vectoren u en v aan te wijzen zodat $a = u + v$, $\lambda b = u - v$. Dit is equivalent met $v = \frac{a - \lambda b}{2}$ en $u = \frac{a + \lambda b}{2}$.
Hiervoor geldt inderdaad

$$(v,u) = \tfrac{1}{4}\{ (a,a) - \lambda^2(b,b) \} = 0.$$

als $a \neq \lambda b$ (maar dan voldoet elke $v \perp u$)

Dus H_v met $v = (a - \lambda b)/2$ voldoet. Maar dan ook $v = a - \lambda b$, aangezien H_v alleen afhangt van de richting van v (wegens (33.5)). Op de volgende stelling is de bovenbedoelde decompositie gebaseerd.

(33.10) Stelling. Zij $a = (a_1, a_2, \dots, a_m)^T$ de eerste kolom van de lin. afbeelding $A : \mathbb{R}^m \to \mathbb{R}^n$. Zij v de vector

$$v = (a_1 \mp \|a\|_2, a_2, \dots, a_m)^T.$$

Dan is $H_v A$ een matrix waarvan de eerste kolom uit nullen bestaat, uitgezonderd het diagonaalelement.

Bewijs: Volgens (33.9) voldoet $v = a \pm \|a\|_2 e_1$, e_1 eerste basisvector.

(33.11) Gevolg: Er is een orthogonale matrix Q en een mxn bovendriehoeksmatrix U zodat $QA = U$.

Bewijs: Stel dat de matrix $A^{(k)}$ reeds van de gedaante

$$A^{(k)} = \begin{array}{cc} \begin{array}{cc} \quad k \quad & \quad n-k \quad \end{array} & \\ \left(\begin{array}{c|c} U^{(k)} & V^{(k)} \\ \hline 0 & W^{(k)} \end{array}\right) & \begin{array}{l} \updownarrow k \\ \updownarrow m-k \end{array} \end{array}$$

is, met $U^{(k)}$ een kxk bovendriehoeks matrix. Men kan een Householder transformatie $H^{(k)}$ toepassen op $W^{(k)}$ die de eerste kolom van $W^{(k)}$ uitgezonderd het diagonaal element nul maakt wegens (33.10).

De matrix

$$S^{(k)} = \begin{array}{cc} \begin{array}{cc} \quad k \quad & \quad n-k \quad \end{array} & \\ \left(\begin{array}{c|c} I & \\ \hline & H^{(k)} \end{array}\right) & \begin{array}{l} k \\ m-k \end{array} \end{array}$$

voert $A^{(k)}$ over in $A^{(k+1)}$. Met de keuze $A^{(0)} = A$ krijgen we zo: $S^{(n-1)} S^{(n-2)} \cdots S^{(2)} S^{(1)} H^{(0)} A = \left(\frac{U^{(n)}}{0}\right)$ waarin $H^{(0)}$ de Householdertransformatie uit (33.10) is.

Of anders geschreven :

$$Q.A = \left(\frac{U^{(n)}}{0}\right)$$

waarin Q als produkt van orthogonale mxm matrices een orthogonale mxm matrix is.

(33.12) Omdat Q orthogonaal is dus afstanden en hoeken behoudt, is de kl.kw. oplossing van Ax = b ook kl.kw. oplossing van Q Ax = Q b , dus van

$$\left(\frac{U^{(n)}}{\emptyset}\right) x = Q\,b \quad , Q\,b \in \mathbb{R}^m.$$

(ga na!)

Een kl.kw. oplossing is gedefiniëerd als die $x \in \mathbb{R}^n$, waarvoor

$$\left\| \left(\frac{U^{(n)}}{0}\right) x - Q\,b \right\|_2$$

minimaal is. Dit betekent dat x oplossing is van het lineaire stelsel

$$U^{(n)}\,x = Q_n . b$$

waarbij Q de matrix is bestaande uit de eerste n rijen van de matrix Q. (Ga na!). Als rang(A) = n is $U^{(n)}$ non-singulier. Als rang(A) < n dan is op zeker moment tijdens het proces de 1e kolom van $W^{(k)}$ nul. Laat die kolom met zijn onbekende gewoon weg, en stel die onbekende later 0.

(33.13) Opgave. Hoeveel AV zijn i.h.a. nodig om $U^{(n)}$ en Q_n.b te berekenen? Hoeveel worteltrekkingen?
Vergelijk dit met het aantal AV nodig om A*A en A*b te berekenen:

(33.14) We leggen nog enige verbanden. Zij A reëel (complex gaat het overigens haast net zo). Zij weer

$$Q\,A = \left(\frac{U^{(n)}}{0}\right) \qquad \text{of} \qquad A = Q^*\left(\frac{U^{(n)}}{0}\right)$$

Er geldt A*A = U*U (ga na), maar ook A*A = C*C, waar C de nxn bovendriehoeksmatrix is volgens de Choleski-decompositie. Nu is C op linksvermenigvuldiging met een matrix S na, uniek bepaald, waarbij S een diagonaalmatrix is met +1 of -1 als diagonaalelement. D.w.z. elke rij van C mag willekeurig met +1 of -1 vermenigvuldigd worden. C is op linksvermenigvuldiging met S na uniek bepaald door de eisen : C*C = A*A en C bovendriehoeks.
U voldoet hieraan, dus is U een decompositiematrix volgens Choleski van A*A en op linksvermenigvuldiging met een S na uniek.
Voor de eerste n kolommen van de matrix Q^T volgt dan, dat zij op rechtsvermenigvuldiging met een matrix S na uniek bepaald zijn d.w.z. de eerste n kolommen opgevat als vectoren in de $\mathbb{R}^m$ mogen naar willekeur met +1 of -1 vermenigvuldigd worden (zelfde of tegengestelde richting). Nu vormen deze kolommen een basis van de n-dimensionale lineaire deelruimte, opgespannen door de kolommen van A. Bovendien worden de kolommen van A uit de eerste n

kolommen van Q^T verkregen door rechtsvermenigvuldiging met een nxn bovendriehoeksmatrix U. Maar dit is ook net het geval als men de eerste n kolommen van Q^T vervangt door de orthogonale basisvectoren die het proces van Gramm-Schmidt oplevert, uitgaande van de kolommen van A. Derhalve zijn de eerste n kolommen van Q^T op een factor $\pm$ 1 na identiek met de basisvectoren volgens Gramm-Schmidt. Bij de oplossing van het probleem van de n^{de}-graads polynoom aanpassing (§ 32) kwamen we uit op een orthonormale basis van polynomen, op het teken na uniek bepaald en identiek met de basis die het proces van Gramm-Schmidt oplevert. Door de identificatie van een polynoom met de vector met als coördinaten de waarde van dit polynoom in een m-tal gegeven punten t_k kan men de n orthonormale polynomen gelijkstellen aan de eerste n kolommen van de matrix Q.

(33.15) Bij de numerieke uitvoering van een Householder transformatie heeft men nog de vrijheid het teken in $a_1 \mp \|a\|_2$ te kiezen (zie (33.10)). Zuiver wiskundig gesproken is de tekenkeuze niet van belang. Er wordt echter in eindige precisie gerekend. Dan is het om numerieke stabiliteit te garanderen noodzakelijk het teken gelijk aan $\underline{\text{sign}}(a_1)$ te kiezen. Voor een bewijs zie Wilkinson, blz. 154 e.v.

(33.16) De methode van de normaalvergelijkingen berekende de kl.kw. oplossing x door het lineaire stelsel $A^*Ax = A^*b$ op te lossen. Perturbaties in A en b zullen dus maximaal een factor $C(A^*A)$ (het conditiegetal van A^*A, in de eucl.norm) vergroot in x doorwerken. Uitgaande van de decompositie verkrijgt men de kl.kw. oplossing als oplossing van $Ux = Q.b$. Men zou verwachten dat perturbaties in A en b nu maximaal met een factor $C(U) = \sqrt{C(A^*A)}$ (ga na!) vergroot in x doorwerken. Helaas is dit niet helemaal waar. In het geval van een groot residu (d.w.z. $\|Ax-b\|$ vergelijkbaar met $\|b\|$) kan men aantonen dat weer de factor $C(A^*A)$ optreedt. Bij klein residu blijken de resultaten verkregen met een decompositie gebruikmakend van Householder transformaties echter veel nauwkeuriger te zijn dan die van de normaalvergelijkingen. Voor details en literatuurlijst zie G.H. Golub: Matrix Decompositions and Statistical Calculations, Tech. Rep no CS 124, 1969, Stanford University.

(33.17) In (33.14) hebben we in feite aangetoond dat de transformatie $QA = U$ ook met het proces van Gramm-Schmidt bepaald kan worden.

Het is echter duidelijk dat zo'n proces nooit de stabiliteit van de Householder-transformaties kan halen. Bij Gramm-Schmidt trekt men immers van een kolom van A een lineaire combinatie van reeds bepaalde basisvectoren af. Neem nu eens aan dat de beschouwde kolom van A bijna afhankelijk is van de vorige. Dan zal de vector, die we na het aftrekken van de lineaire combinatie van reeds bepaalde basisvectoren overhouden, relatief sterk met ruis behept zijn. (ga na; zie ook (25.1)).
Van het loodrecht staan van deze vector op de voorgaande moeten we ons dan geen al te grote voorstelling maken, maar dit betekent dat Q niet erg goed orthogonaal behoeft te zijn, in tegenstelling tot de situatie bij Householder transformaties. Men kan overigens het proces van Gramm-Schmidt modificeren, zodanig dat de orthogonaliteit veel meer behouden blijft. Voor détails zie Golub.

§ 33a Stelsels niet-lineaire vergelijkingen.

(33a.1) Een stelsel niet-lineaire vergelijkingen heeft de vorm

(33a.2)
$$\begin{aligned} f_1(\xi_1,\ldots,\xi_n) &= 0 \\ &\vdots \\ f_n(\xi_1,\ldots,\xi_n) &= 0 \end{aligned}$$

waarin voor elke $i=1,..,n$ $f_i : \mathbb{R}^n \to \mathbb{R}$.
Schrijven we $x = (\xi_1,\ldots,\xi_n)$ en definiëren we
$f(x) = (f_1(\xi_1,\ldots,\xi_n),\ldots,f_n(\xi_1,\ldots,\xi_1))$, dan betekent oplossen van het stelsel (33a.2) het vinden van de wortel(s) van de vergelijking

(33a.3)
$$f(x) = 0$$

met $f : \mathbb{R}^n \to \mathbb{R}^n$ (i.h.a. niet-lineair).

(33a.4) De methoden die wortels van (33a.3) bepalen zijn globaal in 2 groepen te verdelen:
- iteratieve methoden
- minimalisatiemethoden

(33a.5) Iteratief oplossen van de vergelijking $f(x) = 0$ betekent (vgl. §11 e.v.) weer transformatie in $\phi(x) = x$, welke vergelijking men met successieve substitutie zal trachten op te lossen.

(33a.6) Een afbeelding $F : \mathbb{R}^n \to \mathbb{R}^n$ heet differentiëerbaar in punt x indien een lineaire afbeelding $F'(x) : \mathbb{R}^n \to \mathbb{R}^n$ bestaat zodat voor alle h:

$$F(x+h) = F(x) + F'(x)(h) + \varepsilon(x;h)$$

met $\lim\limits_{h\to 0} \dfrac{\|\varepsilon(x;h)\|}{\|h\|} = 0$

$F'(x)$ is dus een matrix t.w. de matrix $\left(\dfrac{\partial F_i}{\partial \xi_j}\right)$ van partiële afgeleiden der coördinaten van F. Men noemt deze matrix ook wel de Jacobiaan.

(33a.7) Als α een wortel is van $\phi(x) = x$ en ϕ is differentieerbaar in α, en we stellen $h_i = x_i - \alpha$ dan geldt blijkbaar

$$\begin{aligned} h_{i+1} &= x_{i+1} - \alpha = \phi(x_i) - \alpha = \phi(\alpha + h_i) - \phi(\alpha) \\ &= \phi'(\alpha)h_i + \varepsilon(h_i) \end{aligned}$$

met $\lim_{h\to 0} \frac{\|\varepsilon(h)\|}{\|h\|} = 0$
zodat $\|\phi'(\alpha)\|$ weer zoiets als de convergentiefactor is.

(33a.8) Voor een willekeurige vergelijking $f(x) = 0$ zoekt men ϕ van de vorm

(33a.9) $$\phi(x) = x-M(x)f(x)$$

waarin voor elke x $M(x)$ een lineaire operator is zodanig dat $\lim_{x\to\alpha} M(x)$ bestaat (α de wortel van $f(x) = 0$).
Als f differentieerbaar is in α dan geldt

$$\begin{aligned}\phi(\alpha+h)-\phi(\alpha) &= h-M(\alpha+h)f(\alpha+h) = \\ &= h-(M(\alpha)+\varepsilon_1(h))(f(\alpha)+f'(\alpha)h+\varepsilon_2(h))\end{aligned}$$

waarin $\varepsilon_1(h)$ een operator en $\varepsilon_2(h)$ een vector is zodat $\lim_{h\to 0}\|\varepsilon_1(h)\| = 0$ en $\lim_{h\to 0} \frac{\|\varepsilon_2(h)\|}{\|h\|} = 0$. Voorts is $f(\alpha) = 0$. Dus

$$\phi(\alpha+h)-\phi(\alpha) = [I-M(\alpha)f'(\alpha)]h+\varepsilon(h)$$

met $\lim_{h\to 0} \frac{\|\varepsilon(h)\|}{\|h\|} = 0$. Dus $\phi'(\alpha) = I-M(\alpha)f'(\alpha)$.

Door te nemen $M(x) = f'(x)^{-1}$ krijgt men het meerdimensionaal analogon van Newton's proces:

(33a.10) $$\phi(x) = x - f'(x)^{-1} f(x)$$

(33a.11) Per iteratiestap zal men dus n^2 afgeleiden moeten berekenen en een matrix inverteren.
Vaak is het bepalen van $f'(x)$ al moeilijk en moet men die eerst benaderen alvorens te inverteren. Dit kan door in de Jacobiaan de differentiaalquotiënten te vervangen door differentiequotiënten en men krijgt het analogon van Koorden-Newton (zie TA §27.4)
Wegens de hoeveelheid werk verbonden aan het bepalen van $f'(x_i)$ en $f'(x_i)^{-1}$ kan het verstandig zijn dit niet bij elke iteratiestap opnieuw te doen, maar slechts af en toe. Werkend met vaste $f'(x_{i0})$ krijgt men weliswaar slechts lineaire convergentie, maar de convergentiefactor is heel klein als x_{i0} dicht bij de wortel ligt.

(33a.12) Voor de keuze van een startwaarde zal men zich moeten beroepen op zekere à priori schattingen omtrent de ligging der wortel(s). Indien deze ontbreken zit er niets anders op dan maar een startpunt te nemen en te hopen dat het iteratie proces naar een wor-

tel convergeert. Met een ander startpunt zal men misschien (hopelijk) een andere wortel vinden etc.
Het is duidelijk dat we op deze manier nooit weten of alle wortels bepaald zijn (indien hun aantal tenminste niet van te voren bekend is). Dit is ook een heel moeilijk probleem.

(33a.13) Bij de minimalisatiemethoden probeert men functionalen $G : \mathbb{R}^n \to \mathbb{R}^n$ te vinden die hun minimum (minima) juist aannemen in de wortel(s) van (33a.3).
Zulke functionalen zijn bv.

$$G(x) = \sum_{i=1}^{n} \left| f_i(\xi_1 \ldots \xi_n) \right|$$

$$G(x) = \sum_{i=1}^{n} \left| f_i(\xi_1, \ldots, \xi_n) \right|^2$$

welke nog de bijzonderheid hebben dat zij in hun (absolute) minima de waarde 0 hebben.

(33a.14) Het vinden van het minimum van G geschiedt veelal met descentmethoden waarbij men, in de i^e iterand x_i aangekomen, een richting r_i zoekt zodat $G(x) < G(x_i)$ voor vectoren $x = x_i + \lambda r_i$ (met voldoend kleine $\lambda > 0$). Dan zoekt men zodanig $\lambda = \lambda_i$ dat in de richting r_i $G(x)$ minimaal wordt, en men vindt $x_{i+1} = x_i + \lambda_i r_i$.

(33a.15) Bij de methode van steepest descent neemt men

(33a.16) $$r_i = - \text{grad } G(x_i)$$

met $\text{grad } G(x) = \begin{pmatrix} \frac{\delta G}{\delta \xi_1} \\ \cdot \\ \cdot \\ \frac{\delta G}{\delta \xi_n} \end{pmatrix}$

Omdat G het sterkst toeneemt in de richting van grad G zal men onder zekere netheidsvoorwaarden mogen stellen dat G het sterkst afneemt in de tegengestelde richting en (28a.16) lijkt inderdaad heel gunstig.

(33a.17) Toch kan het werkelijk convergentiekarakter van de methode van steepest descent erg tegenvallen.
De oorzaak is doorgaans dat - grad G niet de meest gunstige richting is waarin men het minimum moet zoeken.

(33a.18) Bij andere methoden (bv. de geconjugeerde gradiënten methode) zal men vanuit x_i kleine stapjes in een aantal speciale richtingen ondernemen om zo direct mogelijk in het (een)minimum te geraken. Voor nadere details zie bv. Ralston (ch. 8), Kowalik - Osborne.

Eigenwaarden

§ 34 Inleiding

(34.1) Zij A een complexe nxn - matrix.
Een getal $\lambda \in \mathbb{C}$ heet een eigenwaarde van A indien een $x \neq 0$ bestaat zodat $Ax = \lambda x$; x heet in dat geval een eigenvector bij λ.

(34.2) De eigenwaarden van A zijn juist de wortels van het n^e graads polynoom $\Psi(A;\lambda) = \det(A - \lambda I)$, het zgn. karakteristieke polynoom van A.

(34.3) Opgave. Toon aan dat $\Psi(A;\lambda) = \Psi(T^{-1}AT;\lambda)$ voor elke nonsinguliere T.

(34.4) Opgave. Zij A of B nonsingulier.
Toon aan $\Psi(AB;\lambda) = \Psi(BA;\lambda)$. (Via een limietbeschouwing geldt het ook voor A en B singulier. En zelfs geldt als A en B niet vierkant zijn dat $\Psi(AB;\lambda)$ en $\Psi(BA;\lambda)$ maar een factor λ^k verschillen (Wilkinson p54).

(34.5) Opgave. Toon aan dat $\overline{\Psi(A;\lambda)} = \Psi(A^*;\bar{\lambda})$

(34.6) Opgave. Toon aan dat α een wortel is van $a_0x^n + \dots + a_{n-1}x + a_n$ ddan als α een eigenwaarde is van

$$A = \begin{pmatrix} 0 & 1 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ -\frac{a_n}{a_0} & \cdots & \cdots & \cdots & -\frac{a_1}{a_0} \end{pmatrix}$$

Ga ook na dat $(-1)^n a_0 . \Psi(A;\lambda) = a_0\lambda^n + \dots + a_{n-1}\lambda + a_n$. Afgezien van een multiplicatieve factor kan ieder polynoom derhalve optreden als karakteristiek polynoom van een matrix.

(34.7) Opmerking. Men kan $\Psi(A;\lambda)$ zien als formele veelterm en voor λ matrices substitueren.
Stelling van Cayley - Hamilton : $\Psi(A;A) = 0$.

(34.8) Indien $\lambda = \lambda_i$ een eigenwaarde van A is, heeft $\Psi(A;\lambda)$ een deler $(\lambda - \lambda_i)$. Het grootste getal m zodat $(\lambda - \lambda_i)^m \mid \Psi(A;\lambda)$ heet de (algebraïsche) multipliciteit van λ_i. Indien de multipliciteit 1 is noemt men de eigenwaarde enkelvoudig.

Als alle eigenwaarden van de matrix enkelvoudig zijn dan is er een basis van eigenvectoren. Dit laatste kan echter ook wel het geval zijn zonder het eerste (bijv. heeft een symmetrische matrix altijd een basis van eigenvectoren).

(34.9) De eigenwaarden van een matrix A zijn afhankelijk van de elementen van A.
Wijzigt men de elementen van A, dan zullen de eigenwaarden gewoonlijk verschuiven.
De volgende stelling leert dat de eigenwaarden 'continu' afhangen van de elementen van matrix A.

(34.10) Stelling. Er bestaan continue (zelfs analytische) functies $\phi_1,\ldots,\phi_n\colon \mathbb{C}^{n^2}\to\mathbb{C}$ zodanig dat voor elke complexe matrix $A=(a_{ij})$:

$$\Psi(A;\lambda) = (-1)^n.\prod_1^n(\lambda - \phi_i(a_{11}, a_{12},\ldots a_{nn}))$$

Bewijs. De wortels van een polynoom hangen continu van de coefficienten af. Meer precies: er bestaan continue functies $\phi_1,\ldots,\ \phi_n\colon \mathbb{C}^n\to\mathbb{C}$ zodanig dat voor alle $(a_1,\ldots,a_n)\in\ \mathbb{C}^n$:

$$z^n + a_1z^{n-1}+\ldots+a_n = \prod_1^n(z-\ \phi_i(a_1,\ldots,a_n)).$$

Ga na dat de coefficienten van $\Psi\ (A;\lambda)$ continu van de elementen van A afhangen.

§ 35 Localisering van eigenwaarden.

(35.1) De volgende stelling doet een uitspraak over de ligging der eigenwaarden van een matrix.

(35.2) Stelling (Gershgorin's cirkelstelling). Iedere eigenwaarde van $A = (a_{ij})$ ligt in minstens een der gesloten cirkelschijven met middelpunt a_{ii} en straal $\sum_{j\neq i}|a_{ij}|$.
Bewijs. Zij λ e.w. , $x = (x_1,\ldots,x_n)^T$ een e.v. bij λ.
Uit $Ax = \lambda x$ volgt: $(\lambda - a_{rr})x_r = \sum_{j\neq r} a_{rj}\, x_j$ voor alle r.
Dan:

$$|\lambda - a_{rr}|.|x_r| \leq \sum_{j\neq r}|a_{rj}|.|x_j| \quad \text{voor alle } r.$$

Zij i zo dat $|x_i| = \max_r |x_r|$. Dan is $x_i \neq 0$.

$$\Rightarrow\ |\lambda - a_{ii}| \leq \sum_{j\neq i}|a_{ij}|.\frac{|x_j|}{|x_i|} \leq \sum_{j\neq i}|a_{ij}|.$$

(35.3) <u>Opgave</u>. Een bewijs van (35.2) is ook als volgt te verkrijgen.

a) Zij $C = (c_{ij})$.
Toon aan dat $|c_{ii}| > \sum_{j \neq i} |c_{ij}|$ impliceert dat C nonsingulier is.

Vermenigvuldig hiertoe C aan de linkerzijde met zodanige diagonaalmatrix dat het produkt een matrix is met louter enen op de hoofddiagonaal. (Kan dit?) Schrijf dit produkt als I - F (zekere F) en pas (19.24) toe.

b) Pas nu a) toe op de matrix $A - \lambda I$ en bewijs (35.2).

(35.4) <u>Opgave</u>. Ga na dat iedere eigenwaarde van $A = (a_{ij})$ ligt in minstens een der cirkelschijven met middelpunt a_{jj} en straal $\sum_{i \neq j} |a_{ij}|$.

(35.5) De eigenwaarden van A bevinden zich dus in cirkels

$$M_i(A) = \{ z \,\big|\, |z - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \} \qquad i = 1, \ldots, n.$$

Zij heten de <u>Gershgorin - cirkels</u> van A.

(35.6) <u>Opgave</u>. Elke wortel α van $a_0 x^n + \ldots + a_{n-1} x + a_n$ voldoet aan minstens een van de volgende ongelijkheden:

$$\left|\alpha + \frac{a_1}{a_o}\right| \leq 1$$

$$|\alpha| \leq 1 + \max_{i=2,\ldots,n} \left|\frac{a_i}{a_o}\right|$$

(Aanwijzing : (34.6) en (35.4)).

(35.7) <u>Opgave</u>. Elke wortel α van $a_o x^n + \ldots + a_{n-1} x + a_n$ voldoet aan minstens een van de volgende ongelijkheden:

$$|\alpha| \leq 1$$

$$\left|\alpha + \frac{a_1}{a_o}\right| \leq \frac{1}{|a_o|} \cdot \sum_{i=2}^{n} |a_i| .$$

(Aanwijzing : (34.6) en (35.2))

(35.8) De resultaten uit (35.6) en (35.7) maken het mogelijk een uitspraak te doen over de positie der wortels van een polynoom $p(x) = a_o x^n + \ldots + a_{n-1} x + a_n$.
Er zijn nog wel andere wortelschattingen bekend. We noemden reeds

die van Newton voor polynomia met uitsluitend reële wortels. We noemen er nog enkele.
In het nu volgende zullen we steeds aannemen : $a_0 = 1$ (geen beperking).

(35.9) Lemma. Laten α_1 t/m α_n positieve reële getallen zijn met $\sum_1^n \alpha_i \leq 1$. Dan geldt voor elke wortel r van p(x):

$$|r| \leq \max_i \sqrt[i]{\left|\frac{a_i}{\alpha_i}\right|}$$

Bewijs.

$$p(r) = 0 \Rightarrow 1 = \left|\frac{a_1}{r} + \frac{a_2}{r^2} + \ldots + \frac{a_n}{r^n}\right| \leq \sum_1^n \frac{|a_i|}{|r|^i}$$

Dan moet voor minstens één i

$$\frac{|a_i|}{|r|^i} \geq \alpha_i$$

d.w.z. voor minstens één i geldt

$$|r| \leq \sqrt[i]{\left|\frac{a_i}{\alpha_i}\right|}$$

(35.10) Zo krijgt men bv. voor $\alpha_i = 2^{-i}$ als wortelschatting

(35.11) $$|r| \leq 2 \max_i \sqrt[i]{|a_i|}$$

hetgeen vrij gemakkelijk toepasbaar is bij gebruik van een computer met binair getalsysteem ($\sqrt[i]{(m,2^p)}$ met $2^{q-1} \leq |m| < 2^q$ majoreert men dan niet al te slecht door $2^{1+\text{entier}((p+q)/i)}$).

(35.12) Opgave. Ga na wat de schatting uit (35. 9) wordt voor $\alpha_i = |a_i| / \sum_1^n |a_j|$.
Geef het max. expliciet aan. Onderscheidt daarbij tussen $\sum_1^n |a_j| \leq 1$ resp. ≥ 1.

(35.13) Opgave. Ga na dat (35.9) ook toepasbaar is met

$$\alpha_i = |a_i| \Big/ \left(\sum_1^n \sqrt[j]{|a_j|}\right)^i$$

Schrijf hiertoe $c_j = \sqrt[j]{|a_j|}$ en bezie $\sum_1^n \left(\frac{c_i}{\sum_1^n c_j}\right)^i$.

Wat wordt de schatting uit (35.9) voor deze α_i.

(35.14) Het is moeilijk te zeggen welke van de gegeven schattingen de beste is; soms is de een, soms de ander iets beter. Men kan echter aantonen dat (35.11) nooit meer dan een factor 2n te groot is.

(35.15) We merken op dat niet alle schattingen op dezelfde wijze reageren op schaalverandering.
Bij schaalverandering gaat men naast $p(x) = x^n + a_1x^{n-1} + \dots$ bezien het polynoom met k-keer zo grote wortels.
Ga na dat dit het polynoom

$$q(x) = x^n + k\,a_1x^{n-1} + \dots + k^{n-1}a_{n-1}x + k^na_n.$$

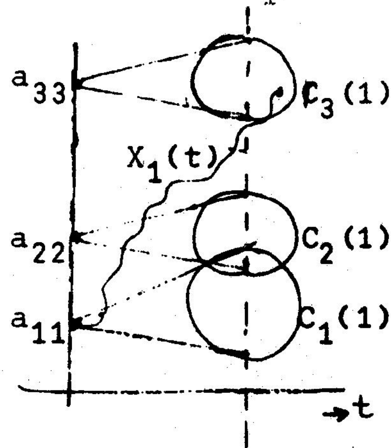
is.
Als de schatting voor de wortels van $q(x)$ k x zo groot is als de overeenkomstige schatting voor de wortels van $p(x)$ noemt men de wortelschatting <u>homogeen</u>, anders <u>inhomogeen</u>.

(35.16) <u>Opgave</u>.
a) (35.11) en (35.13) zijn homogene schattingen.
b) (35.12), (35.6) en (35.7) zijn inhomogene schattingen.
Ga voor verschillende polynomia met bekende wortels (bijv. 1,1,1 en $10^6,10^6,10^6$) eens na wat er uitkomt.

(35.17) We formuleren nog een tweede stelling, die een interessante eigenschap van de Gershgorin-cirkels geeft.

(35.18) <u>Stelling</u> (<u>Gershgorin's 2^e cirkelstelling</u>). Indien k Gershgorin-cirkels van A geïsoleerd liggen van de overige, dan bevatten deze k cirkels tezamen precies k eigenwaarden van A, elke eigenwaarde geteld naar zijn multipliciteit.
<u>Bewijs</u>:
Schrijf $A = D + B$, D een diagonaalmatrix met $d_{ii}=a_{ii}(i=1,\dots,n)$.
Zij $A(t) = D + tB$ voor $t \in [0,1]$, dus $A(o)=D$ en $A(1)=A$. Zij $C_i(t) = \{z \mid |z-a_{ii}| < t \sum_{j\neq i} |a_{ij}|\}(i=1,\dots,n)$
Dan zijn de $C_i(t)$ de Gershgorin-cirkels van $A(t)$.
Wegens (34.10) bestaan continue functies $X_1,\dots,X_n: \mathbb{R} \to \mathbb{C}$ zodat $X_1(t),\dots,X_n(t)$ de eigenwaarden van $A(t)$ zijn.
Na eventueel vernummeren der functies X_i geldt voor alle i: $X_i(o) = a_{ii}$.
Stel nu dat k Gersh gorin-cirkels van A geïsoleerd liggen van de overige. Het is geen restrictie te onderstellen dat dit $C_1(1),\dots,C_k(1)$ zijn. Evident liggen dan voor alle $t \in [0,1]$ $C_1(t),\dots,C_k(t)$ geïsoleerd liggen van de overige $C_i(t)$.

Als nu b.v. $X_1(1) \in C_{k+1}(1)$, dan moet wegens $X_1(o)=a_{11}$



een t bestaan zodat $X_1(t)$ in geen enkele $C_i(1)$ ligt, dus zeker in geen enkele $C_i(t)$. (zie b.v. nevenstaande figuur en (35.19)). Dit is eén tegenspraak! Etc.
Men ziet in dat $X_1(1),\ldots,X_k(1)$ noodzakelijk in de vereniging der $C_1(1),\ldots, C_k(1)$ moeten liggen.

(35.19) In de bovenstaande figuur is het geval geschetst dat van reele 3 × 3 A twee Gershgorin-cirkels geïsoleerd liggen van de overige.
De aangegeven cirkels moet men loodrecht op het vlak van tekening denken.
Beziet men de Gershgorin-cirkels met middelpunt a_{ii} als t van 0 naar 1 loopt, dan vormen zich kegels, waarvan de projecties zijn aangegeven.
De functie $X_1(t)$ zoals getekend kan nooit het verloop van een eigenwaarde zijn. Die moet te allen tijde binnen een Gershgorin-cirkel liggen, hetgeen betekent dat $X_1(t)$ op elk moment binnen een kegel moet liggen.

(35.20) Opgave. Ga na hoe het argument uit (35.19) in een willekeurige situatie te voeren is.

(35.21) Opgave. Zij $A= D + \epsilon B$, D een diagonaalmatrix met $d_{ii} = a_{ii}$.
Stel voor alle $i= 2,\ldots,n$: $a_{ii} \neq a_{11}$; zij $\delta = \min_{i\geq 2}|a_{11}-a_{ii}|$.
Toon aan dat voor alle ϵ met $|\epsilon| < \frac{\delta}{2\|B\|_\infty}$ binnen de Gershgorin-cirkel rond a_{11} precies één enkelvoudige eigenwaarde van A ligt.

§36 Perturbatie van eigenwaarden.

(36.1) Verstoringen in de matrix A beïnvloeden de ligging der eigenwaarden en eigenvectoren.
We zullen dit effect onderzoeken voor het geval dat A een diagonaliseerbare matrix is.

(36.2) Laat $T^{-1} AT = D$, met T nonsingulier, $D = \begin{pmatrix} \lambda_1 & & \Theta \\ & \ddots & \\ \Theta & & \lambda_n \end{pmatrix}$ een diagonaalmatrix waarvan de diagonaalelementen dus juist de eigenwaarden van A zijn.
We bezien eerst het effect van perturbaties ε B.
Zij $T^{-1} BT = (\beta_{ij})$.

(36.3) Stelling. Zij λ_i enkelvoudige eigenwaarde van een diagonaliseerbare matrix A.
Dan heeft $A + \varepsilon B$ voor voldoend kleine ε precies één, eveneens enkelvoudige eigenwaarde μ_i waarvoor geldt:

$$|\mu_i - \lambda_i - \varepsilon\, \beta_{ii}| \leq \kappa .\ |\varepsilon|^2 \sum_{j \neq i} |\beta_{ij}|$$

met zekere constante κ, die onafhankelyk van ε is.

Bewijs.
De eigenwaarden van $A + \varepsilon B$ zijn gelijk aan die van
$T^{-1}(A + \varepsilon B)T = D + \varepsilon T^{-1} BT$, en dus ook gelijk aan die van
$P^{-1}(D + \varepsilon T^{-1} BT)P = D + \varepsilon\, P^{-1}T^{-1} B\,T\,P$ voor willekeurige nonsinguliere diagonaalmatrix P.
Kies nu in P het i^e diagonaalelement gelijk aan $(\kappa\varepsilon)^{-1}$, en de overige diagonaalelementen 1.
$P^{-1}T^{-1} B\,T\,P$ ontstaat uit $T^{-1} BT = (\beta_{ij})$ door de i^e kolom te vermenigvuldigen met $(\kappa\varepsilon)^{-1}$ en de i^e-rij te vermenigvuldigen met $\kappa\varepsilon$.
Het element op de plaats (i,i) verandert dus niet.
Bezie de Gershgorin - cirkels van $D + \varepsilon P^{-1}T^{-1} B\,T\,P$.
We weten uit de enkelvoudigheid dat λ_i verschillend is van alle overige elementen op de hoofddiagonaal van D. Ofwel: de Gershgorincirkel $M_i(D)$ (straal = 0) ligt geïsoleerd van de overige Gershgorincirkels.
Maar dan zal voor ε met $|\varepsilon| <$ zekere τ de Gershgorincirkel van $D + \varepsilon P^{-1}T^{-1} B\,T\,P$ met middelpunt $\lambda_i + \varepsilon\beta_{ii}$ en straal $\kappa|\varepsilon|^2 \sum_{j\neq i} |\beta_{ij}|$ ook nog geïsoleerd liggen van de overige, mits κ voldoende groot en τ voldoende klein.

Wegens (35.6) zal deze cirkel juist één, enkelvoudige eigenwaarde van $D + \varepsilon P^{-1}T^{-1}B\ T\ P$, dus van $A + \varepsilon B$ bezitten.

(36.4) Uit (36.3) ziet men dat een <u>enkelvoudige</u> eigenwaarde λ_i van een diagonaliseerbare matrix A in eerste benadering verstoord wordt met een bedrag $\varepsilon\beta_{ii}$.

(36.5) <u>Opgave</u>. Toon aan dat men in het bovenstaande bewijs kan nemen:

$$\kappa > \max_{r \neq i} \frac{\sum_{j=1}^{n} (|\beta_{rj}| + |\beta_{ij}|)}{|\lambda_r - \lambda_i|}$$

$$\tau = \frac{1}{\kappa}$$

bij diagonaliseerbare matrices

(36.6) Gaat men het effect van perturbaties εB op een niet-enkelvoudige zeg m-voudige eigenwaarde λ na, dan vindt men analoog aan het bewijs van (36.3) m Gershgorincirkels die voor voldoend kleine ε geïsoleerd liggen van de overige.

Dit geeft niet veel meer informatie dan dat er m eigenwaarden μ van $A + \varepsilon B$ bestaan zodat :

$$|\mu - \lambda| \leq \varepsilon .\ \|T^{-1}\ BT\|_{\infty}. \qquad \text{(Ga na).}$$

De volgende stelling leert dat algemener geldt:

36.7) <u>Stelling</u>. (Bauer - Fike). Zij A een diagonaliseerbare matrix, $T^{-1}\ AT = D$.

Zij μ een eigenwaarde van $A + B$.

Dan geldt: $\min_i |\mu - \lambda_i| \leq \|T^{-1}\ BT\|_p$ voor alle $p \in [1, \infty]$.

<u>Bewijs</u>. We mogen aannemen dat $\min_i|\mu - \lambda_i| \neq 0$, d.w.z. dat μ geen eigenwaarde van A is (anders was er niets te bewijzen).

$T^{-1}(A - \mu I)T = D - \mu I$ is dus nonsingulier.

$T^{-1}(A + B - \mu I)T = D - \mu I + T^{-1}\ B\ T$ is singulier zodat ook $(D - \mu I)^{-1}\ (D - \mu I + T^{-1}\ BT) = I + (D - \mu I)^{-1}T^{-1}BT$ singulier is.

Wegens (19.24) moet dan $\|(D - \mu I)^{-1}T^{-1}BT\|_p \geq 1$ voor $1 \leq p \leq \infty$.

Ga na dat $\|(D - \mu I)^{-1}T^{-1}BT\|_p \leq \frac{1}{\min_i|\mu-\lambda_i|}\ \|T^{-1}BT\|_p$.

(36.8) Voor symmetrische A volgt met $p = 2$ dat binnen een afstand $\|B\|_2$ van een eigenwaarde μ van de geperturbeerde matrix $A + B$ minstens één eigenwaarde van A te vinden is. Door deze stelling wordt evenwel niet uitgesloten dat alle eigenwaarden van de geperturbeerde matrix in de buurt van één bepaalde e.w. van A liggen.

Als A en B beide symmetrisch zijn kan dit laatste niet optreden en is een veel preciesere relatering van de eigenwaarden mogelijk (zie echter 36.19):

(36.9) Stelling. Zij A symmetrisch met eigenwaarden $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$.
Zij A + B symmetrisch met eigenwaarden $\mu_1 \leq \mu_2 \leq \dots \leq \mu_n$.
Dan $|\mu_i - \lambda_i| \leq \|B\|_2$, $i = 1, \dots n$.
(Voor een bewijs van (36.9) wordt de geïnteresseerde lezer verwezen naar Wilkinson - The Algebraic Eigen value Problem, pg 99-102.)

(36.10) Opgave. Zij $A = (a_{ij})$ een symmetrische matrix waarvan de coëfficiënten niet alle computergetallen zijn in de zin van § 4 . De k^e eigenwaarde van A en de k^e eigenwaarde van de computer-representatie van A verschillen hoogstens $n. \|A\|_m. \xi$, waarin $\|A\|_m = \max_{i,j} |a_{ij}|$.

(36.11) Opgave. Zij $A = (a_{ij})$ een symmetrische matrix met eigenwaarden $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$.
Zonder beperking mogen we aannemen: $a_{11} \leq a_{22} \leq \dots \leq a_{nn}$.
Dan geldt: $|\lambda_i - a_{ii}| \leq \max_{i} \sum_{j \neq i} |a_{ij}|$.
(vgl. (35.2)).

(36.12) Zij A symmetrisch, B een symmetrische perturbatie van A.
Zo'n perturbatie kan bv. het gevolg zijn van afrondfouten in de coëfficiënten van A of optreden als de coëfficiënten empirisch bv. door meting bepaald worden.
Voor het verschil $\Delta\lambda_k = \mu_k - \lambda_k$ (μ en λ als in (36.9)) geldt:

(36.13) $$|\Delta\lambda_k| \leq \|B\|_2 = \rho(B) \leq n. \|B\|_m$$
zodat wegens $\rho(A) = \|A\|_2 \geq \max |a_{ij}| = \|A\|_m$ geldt:

(36.14) $$\left|\frac{\Delta\lambda_k}{\rho(A)}\right| \leq n. \frac{\|B\|_m}{\|A\|_m} .$$

(36.15) De verstoring in de k^e eigenwaarde relatief t.o.v. de absoluut grootste is dus hoogstens n-keer zo groot als de maximale verstoring in de matrixelementen relatief t.o.v. het absoluut grootste matrixelement. Dit sluit natuurlijk niet uit dat $\frac{|\Delta\lambda_k|}{|\lambda_k|}$ best groot kan zijn. Men lette bv. op de eigenwaarden die in abs. waarde klein zijn t.o.v. $\rho(A)$.

(36.16) Voor de verstoring in de k^e eigenwaarde tengevolge van afrondfouten in de coëfficiënten van A geldt dus (zie ook (36.10)):

$$\frac{|\Delta\lambda_k|}{\rho(A)} \leq n.\ \bar{\xi}$$

zodat men mag stellen dat een symmetrische matrix zijn eigenwaarden op een uitstekende wijze representeert.

(36.17) Bij niet-symmetrische matrices kan de situatie veel ongunstiger zijn.

(36.18) <u>Voorbeeld</u>:

$$A = \begin{pmatrix} 20 & 20 & & & \\ & 19 & \ddots & & \Theta \\ & & \ddots & \ddots & \\ & \Theta & & \ddots & 20 \\ & & & & 1 \end{pmatrix}$$

$$B = \begin{pmatrix} 0 & & & \\ \vdots & & \epsilon & \\ \vdots & & & \\ 1 & \cdots & \cdots & 0 \end{pmatrix} \qquad (20 \times 20)$$

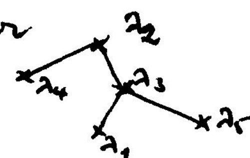
De eigenwaarden van A zijn $1 < 2 < \ldots < 20$.

$$\Psi(A + \varepsilon B;\lambda) = \prod_{i=1}^{20} (i-\lambda) - 20^{19}\varepsilon = \Psi(A;\lambda) - 20^{19}\varepsilon$$

Omdat op $[1,20]$: $|\Psi(A;\lambda)| < 20!$ (ruime bovengrens) zal voor $\varepsilon > \frac{20!}{20^{19}} \approx 4,5 \cdot 10^{-7}$ het polynoom $\Psi(A + \varepsilon B;\lambda)$ nog slechts 2 reële wortels hebben, de overige zijn complex! Voor $\varepsilon \leq -\frac{20!}{20^{19}}$ geen enkele wortel reëel.

(36.19) <u>Stelling</u> Zij A diagonaliseerbaar. Zij B willekeurig. Dan bestaat er een nummering van de eigenwaarden λ_i van A en μ_i van A+B zodat $|\lambda_i - \mu_i| \leq (2n-1)\|T^{-1}BT\|_\infty$, T de diagonaliserende transformatie van A.

<u>Bewijs</u> Teken de λ_i in het complexe vlak. Verbind elk tweetal der λ_i die in absolute waarde minder dan $2\|T^{-1}BT\|$ van elkaar liggen. Beschouw een samenhangend stuk van de zo ontstane graaf. De Gershgorincirkels hiervan liggen geisoleerd van alle andere, en bevatten dus evenveel μ's als λ's. Nu kan men bij elke nummering der λ_i een geschikte nummering der μ_i vinden.



§ 37 Perturbatie van eigenvectoren.

(37.1) Zij A weer een diagonaliseerbare matrix en stel

$$T^{-1} AT = D = \begin{pmatrix} \lambda_1 & & \theta \\ & \ddots & \\ \theta & & \lambda_n \end{pmatrix}$$

Zij λ_i een enkelvoudige eigenwaarde van A.
Het is geen beperking i = 1 te nemen.

(37.2) Uit (36.3) volgt dat onder alle eigenwaarden van $A + \varepsilon B$ met $|\varepsilon| < \tau$, $\tau > 0$ voldoende klein, er precies één, zeg $\mu(\varepsilon)$, is die het dichtst bij λ_1 ligt.
Deze $\mu(\varepsilon)$ is bovendien enkelvoudig zodat de bijbehorende eigenruimte evenals die van λ_1 1-dimensionaal is.

(37.3) Zij $v(\varepsilon)$ een eigenvector van $A + \varepsilon B$ bij eigenwaarde $\mu(\varepsilon)$, v een eigenvector van A bij eigenwaarde λ_1 ($v(\varepsilon)$ en v zijn op scalaire factoren na uniek bepaald).

(37.4) We willen nagaan hoe goed de richting van $v(\varepsilon)$ de richting van v benadert.

(37.5) Gebruikmakend van de diagonaliseerbaarheid gaan we natuurlijk kijken naar $z = T^{-1}v$ en $z(\varepsilon) = T^{-1} v(\varepsilon)$ die eigenvectoren van $T^{-1} AT = D$ resp. $T^{-1}(A + \varepsilon B)T = D + \varepsilon(\beta_{ij})$ zijn.

(37.6) Lemma. Zij C een singuliere n x n matrix die gepartitioneerd is als $\begin{pmatrix} P & Q \\ R & S \end{pmatrix}$, P een k x k matrix, S een (n-k) x (n-k) matrix.
Zij $x = (x_1, \ldots, x_n)^T$ een vector waarvoor $Cx = 0$.
Partitioneer x als $\begin{pmatrix} u \\ v \end{pmatrix}$, waarin u staat voor de eerste k coördinaten, v voor de overige (n-k) coördinaten van x.
Zij S nonsingulier.
Dan geldt $v = -S^{-1} R u$.
Bewijs. $Ru + Sv = 0$.

(37.7) We passen dit lemma toe voor k = 1 op de matrix
$C = D + \varepsilon(\beta_{ij}) - \mu(\varepsilon) I$.

(37.8) Lemma. Voor voldoend kleine ε is de (n-1) x (n-1) rechtsondermatrix S van $C = D + \varepsilon(\beta_{ij}) - \mu(\varepsilon) I$ nonsingulier.
Bewijs. Definiëer $\kappa = \min_{i\neq 1} |\lambda_1 - \lambda_j|$

Zij $\tau_1 > 0$ zodanig dat voor alle ε met $|\varepsilon| < |\tau_1|$:
$\min_{i\neq 1} |\mu(\varepsilon) - \lambda_i| \geq \frac{1}{2}\kappa$.

Zij $\tau_2 = \dfrac{\kappa}{2 \max_{i\neq 1} \sum_{j=1}^{n} |\beta_{ij}|} = \dfrac{\kappa}{2 . \|T^{-1}BT\|_\infty}$

Ga na dat voor $|\varepsilon| < \min(\tau, \tau_1, \tau_2)$ de matrix S voldoet aan de condities van (35.3a)

(37.9) Derhalve kunnen we voor de eigenvector $z(\varepsilon)$ van $D + \varepsilon(\beta_{ij})$ stellen dat

(37.10) $$z(\varepsilon) = \begin{pmatrix} u \\ v \end{pmatrix}$$

met $v = -S^{-1}R\, u$.

(37.11) Hierin is u in feite de eerste coördinaat van $z(\varepsilon)$ die we gevoeglijk gelijk aan 1 mogen nemen.
Dan is

(37.12) $$z(\varepsilon) = \begin{pmatrix} 1 \\ v \end{pmatrix}$$

met

(37.13) $v = -S^{-1} (\varepsilon\beta_{21}, \ldots, \varepsilon\beta_{n1})^T$

(37.14) Men ziet dat voor voldoend kleine ε de vector $z(\varepsilon)$ continu afhangt van ε en dat in het bijzonder $z(\varepsilon)$ steeds meer de richting van z krijgt.

(37.15) Schrijf $S = D' + \varepsilon B'$ met

$$D' = \begin{pmatrix} \lambda_2 - \mu(\varepsilon) & & \Theta \\ & \ddots & \\ \Theta & & \lambda_n - \mu(\varepsilon) \end{pmatrix}$$

$$B' = \begin{pmatrix} \beta_{22} & \cdots & \beta_{2n} \\ | & & | \\ \beta_{n2} & \cdots & \beta_{nn} \end{pmatrix}$$

Dan is wegens (20.12) en wegens $\mu(\varepsilon) = \lambda_1 + O(\varepsilon)$

$$S^{-1} = \begin{pmatrix} \frac{1}{\lambda_2 - \mu(\varepsilon)} & & \\ & \ddots & \\ & & \frac{1}{\lambda_n - \mu(\varepsilon)} \end{pmatrix} + O(\varepsilon) = \begin{pmatrix} \frac{1}{\lambda_2 - \lambda_1} & & \\ & \ddots & \\ & & \frac{1}{\lambda_n - \lambda_1} \end{pmatrix} + O(\varepsilon)$$

zodat

$$v = \varepsilon \begin{pmatrix} \frac{\beta_{21}}{\lambda_1 - \lambda_2} \\ \vdots \\ \frac{\beta_{n1}}{\lambda_1 - \lambda_n} \end{pmatrix} + O(\varepsilon^2)$$

(37.21) Voor voldoend kleine ε mogen we derhalve stellen dat in eerste orde benadering:

(37.22) $$z(\varepsilon) \approx \left(1, \frac{\varepsilon\beta_{21}}{\lambda_1 - \lambda_2}, \ldots, \frac{\varepsilon\beta_{n1}}{\lambda_1 - \lambda_n}\right)^T$$

Terugtransformeren met T geeft dan de eerste orde benadering voor $v(\varepsilon)$.

(37.23) Laten we nu aannemen dat T unitair is.
Ga na dat dit juist optreedt in het geval A een normale matrix is. (vgl. (19.10) en (19.12)).

Omdat T en dus T^{-1} inprodukten behoudt, is de hoek $\phi(\varepsilon)$ tussen $v(\varepsilon)$ en v gelijk aan die tussen $z(\varepsilon)$ en z zodat:

(37.24)
$$\operatorname{tg} \phi(\varepsilon) \approx |\varepsilon| . \left(\sum_{j=2}^{n} \frac{|\beta_{j1}|^2}{|\lambda_I - \lambda_j|^2} \right)^{\frac{1}{2}} \leq \frac{1}{\kappa} \left(\sum_{j=2}^{n} |\beta_{j1}|^2 \right)^{\frac{1}{2}} . \quad |\varepsilon|$$

We hebben nu:

(37.25) Stelling. Zij λ_i een enkelvoudige eigenwaarde van een normale matrix A, $\kappa = \min_{j \neq i} |\lambda_i - \lambda_j|$
Zij ε voldoende klein.
Onder alle eigenwaarden van $A + \varepsilon B$ is er precies één ($\mu_i(\varepsilon)$) die het dichtst bij λ_i ligt.
Deze $\mu_i(\varepsilon)$ is enkelvoudig en voor de hoek $\phi(\varepsilon)$ tussen een eigenvector bij $\mu_i(\varepsilon)$ en een eigenvector bij λ_i geldt:
$$|\phi(\varepsilon)| \leq \frac{\|B\|_2}{\kappa} |\varepsilon|$$

(37.26) De perturbatie van eigenvectoren bij een meervoudige eigenwaarde is een aanmerkelijk moeilijker aangelegenheid.
Een eigenvector is dan niet meer op een scalaire factor na uniek bepaald.
Men kan hoogstens zeggen dat de eigenruimte bij de gepertubeerde eigenwaarde steeds dichter "nadert" tot de oorspronkelijke eigenruimte indien $\varepsilon \to 0$.

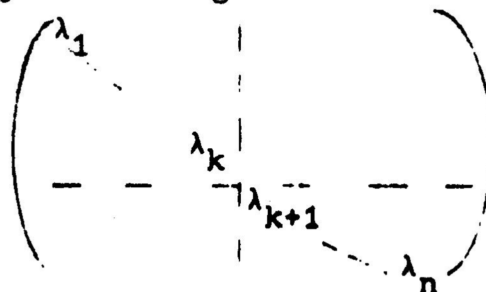
(37.27) Voor symmetrische perturbaties $A + \varepsilon B$ van symmetrische matrices A kunnen we dit nog nader analyseren.
De volgende stelling is ook toepasbaar op het geval dat een aantal eigenwaarden van A dicht bijeen liggen of samenvallen.

(37.28) Stelling. Laten A en B symmetrische matrices zijn. Stel A heeft eigenwaarden $\lambda_1, \ldots, \lambda_n$ (in een of andere volgorde).
Zij V de eigenruimte opgespannen door de eigenvectoren bij $\lambda_1, \ldots, \lambda_k$ en zij $V^{\perp}$ het orthoplement van V.
Zij λ de geperturbeerde waarde van een der $\lambda_1, \ldots, \lambda_k$ zodat dus $|\lambda - \lambda_i| \leq |\varepsilon| \|B\|_2$ (zekere $i=1,\ldots k$) en zij x een bijbehorende eigenvector van $A + \varepsilon B$.
Schrijf $x = x^{\parallel} + x^{\perp}$, $x^{\parallel} \in V$, $x^{\perp} \in V^{\perp}$

Zij $\rho = \min_{j>k,\ l\leq k} |\lambda_j - \lambda_l| - 2\varepsilon \|B\|_2 > 0$.

Dan is $\|x^{\perp}\|_2 \leq \frac{|\varepsilon|}{\rho} \|B\|_2 \ \|x^{\parallel}\|_2$.

Bewijs. De stelling is basis-onafhankelijk en we mogen dus veronderstellen dat A nevenstaande gedaante heeft. Partitioneer $A + \varepsilon B - \lambda I$ evenzo: $\begin{pmatrix} P & Q \\ R & S \end{pmatrix}$. Dan zijn alle e.w. van S in abs. waarde $\geq \rho$, dus $\|S^{-1}\|_2 \leq \frac{1}{\rho}$. Voorts geldt



$$\begin{pmatrix} x_{k+1} \\ \vdots \\ x_n \end{pmatrix} = -S^{-1}R \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix}$$

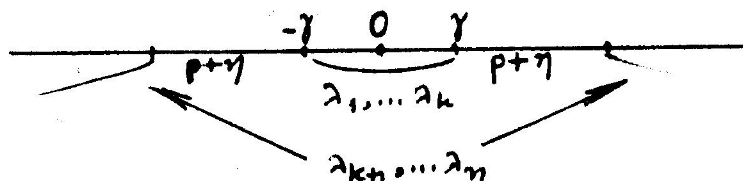
waaruit wegens $\|R\|_2 \leq |\varepsilon| \ \|B\|_2$ het beweerde volgt.

(37.29) Bijgevolg, als de eigenwaarden $\lambda_1, \ldots \lambda_k$ van A weinig van elkaar verschillen of eventueel samenvallen, maar behoorlijk gescheiden liggen van de overige eigenwaarden, dan heeft de bijbehorende eigenruimte van $A + \varepsilon B$ een basis waarvan alle vectoren hoogstens een hoek $\approx \frac{\varepsilon}{\rho} \|B\|_2$ met de eigenruimte bij $\lambda_1, \ldots, \lambda_k$ van A maken.

Een nog wat scherper resultaat vindt men in het zeer fraaie (maar moeilijke) artikel van C. Davis en W.M. Kahan: The rotation of eigenvectors by a perturbation III. SIAM. J. Num. Anal. 7 (1970) 1-46:

(37.30) Stelling. Laten A en B symmetrische matrices zijn. Zij V eigenruimte behorend bij de eigenwaarden $\lambda_1, \ldots, \lambda_k$ van A (in een of andere nummering). Laten $\lambda_1, \ldots, \lambda_k$ in een interval (a, b) liggen en $\lambda_{k+1}, \ldots, \lambda_n$ er buiten of andersom. Zij $\rho = \min_{j>k,\ l\leq k} |\lambda_j - \lambda_l| - \varepsilon \|B\|_2 > 0$. Zij W eigenruimte behorend bij de eigenwaarden van $A + \varepsilon B$ die in de zin van (36,9) bij $\lambda_1, \ldots, \lambda_k$ horen. Zij θ de maximale hoek tussen de vectoren van W en hun projectie op V. Dan is $\sin(\theta) \leq \frac{\varepsilon}{\rho} \|B\|_2$.

Bewijs (eenvoudiger dan bij Davis-Kahan). $A + \alpha I$ heeft voor willekeurige α hetzelfde eigensysteem als A. Het is daarom geen beperking aan te nemen dat hetzij $\lambda_1, \ldots, \lambda_k$ een segment $[-\gamma, \gamma]$ opspannen, hetzij $\lambda_{k+1}, \ldots \lambda_n$. We nemen het eerste aan. Het andere gaat analoog. We krijgen dan de situatie van de tekening, waarin nog $\eta = \varepsilon \|B\|_2$



A beeldt $V^\perp$ in zichzelf af, en daarop geldt $\|Ay\| \geq (\gamma+\rho+\eta)\|y\|$ (zie (20.6)), zodat deze afbeelding op is.

$A+\varepsilon B$ beeldt W in zichzelf af, en daarop geldt $\|(A+\varepsilon B)y\| \leq (\gamma+\eta)\|y\|$ (zie (20.6) en (36.9)).

Zij nu u, w een vectorpaar, $u \in V^\perp$, $w \in W$, $\|u\| = \|w\| = 1$, waarvoor (u, w) (inproduct) maximaal is. Er is een $v \in V^\perp$, $\|v\| = 1$, zodat $Av = cu$, $c > 0$, en dus $c \geq \gamma+\rho+\eta$. Nu geldt enerzijds

$$|(Av, w) - ((A+\varepsilon B)v, w)| = \varepsilon|(Bv, w)| \leq \varepsilon\|B\| = \eta$$

Anderzijds geldt

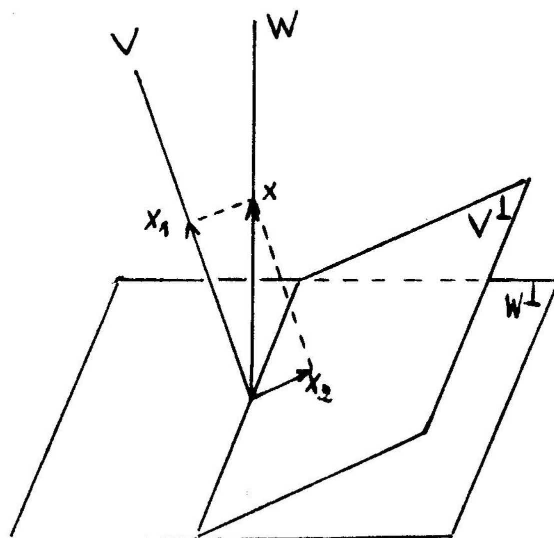
$$(Av, w) - ((A+\varepsilon B)v, w) = (Av, w) - (v, (A+\varepsilon B)w) \geq$$

$$(\gamma+\rho+\eta)(u, w) - (\gamma+\eta)(u, w) = \rho(u, w)$$

zodat $(u, w) \leq \dfrac{\eta}{\rho}$.

Zij nu $x \in W$, x willekeurig. Schrijf $x = x_1 + x_2$, $x_1 \in V$, $x_2 \in V^\perp$. Dan is $\sin(\angle(x, x_1)) = \cos(\angle(x, x_2)) \leq (u, w) \leq \eta/\rho$.

Dus $\sin(\theta) \leq \dfrac{\varepsilon}{\rho}\|B\|_2$.



§ 38 Methoden voor de bepaling van eigenwaarden en eigenvectoren; inleiding.

Er zijn 3 typen van methoden te onderscheiden:

- Methoden waarbij het karakteristieke polynoom $\Psi(A;\lambda)$ expliciet bepaald wordt. (De eigenwaarden vindt men dan als wortels van dit polynoom).
- Methoden die berusten op rechtstreekse bepaling van de nulpunten van de functie $\Psi(A;\lambda) = \det(A-\lambda I)$, maar daartoe A eerst transformeren tot een gedaante waarvan gemakkelijker de determinant te berekenen is. (Methoden van Householder, Givens)
- Methoden die berusten op een rechtstreekse iteratie met matrices (machtsmethode, Jacobi's methode, LU en QR - algorithmen)

§ 39 Methoden die het karakteristieke polynoom expliciet bepalen.

(39.1) Een uitgebreide behandeling van deze methoden vindt men in Faddeev & Faddeeva (Ch IV).
Zie ook bv. P.A. White Jsiam 6 (1958) 393-437.

(39.2) De methode van Krylov maakt gebruik van het feit dat $\Psi(A;A) = 0$ (zie 34.7)).
Stellen we $\Psi(A;\lambda) = \sum_0^n a_k \lambda^k$ met $a_n = 1$ dan geldt voor willekeurige x:

$$\Psi(A;A)x = (\sum_0^n a_k A^k)x = 0.$$

Genereer nu de rij $\{x_i\}_{i=0,\ldots,n}$ met

$$x_0 = x$$
$$x_i = Ax_{i-1} \qquad i \geq 1$$

dan geldt:

$$\sum_0^n a_k x_k = 0$$

en dit geeft in coördinaten een stelsel van n lineaire vergelijkingen voor de coëfficienten $\{a_i\}_{i=0,\ldots,n-1}$.
Ga na dat het bepalen der coëfficiënten ons op deze wijze $\approx \frac{4}{3}n^3$ AV kost.

(39.3) Opgave. Ga na dat het stelsel singulier is als A twee Jordankastjes heeft met gelijke eigenwaarden.

(39.4) Andere rekenschema's zijn afkomstig van Leverrier, Danilevski e.a. en berusten er bv. op dat spoor $(A^k) = \Sigma\lambda_i^k$, (en hieruit kan men de coëfficiënten van het eigenwaardenpolynoom zonder veel moeite bepalen).

(39.5) Vanzelfsprekend maakt men bij het bepalen van de coëfficiënten van het eigenwaardenpolynoom afrondfouten, soms zelfs heel ernstige (vgl. (39.3)). Maar zelfs als men de coëfficiënten heel nauwkeurig te pakken krijgt, kunnen de wortels nog ernstig verstoord zijn.

(39.6) Als afschrikwekkend voorbeeld van de invloed van kleine perturbaties op de ligging der wortels moge dienen het karakteristiek polynoom van de matrix A uit (36.18):

$$\Psi(\lambda) = \prod_{i=1}^{20} (\lambda - i) = \lambda^{20} + a_{19}\lambda^{19} + \dots + a_0 .$$

Aangezien de wortels van een polynoom analytisch van de coëfficiënten afhangen geldt voor een (enkelvoudige) wortel p van Ψ:

(39.7)
$$\Delta p \approx - \frac{p^k}{\Psi'(p)} \Delta a_k = (-1)^{p+1} \frac{p^k}{(p-1)!\ (20-p)!} \Delta a_k$$

Voor p = 16 en k = 15 : $\Delta p \approx 4.10^4\ \Delta a_k$.

Stel dat a_{15} ($\approx 10^{10}$) in de rekenautomaat waarmee men de oplossing wenst te bepalen niet exact voorstelbaar is.

De afrondfout $a_{15}\ \xi$ (vgl. (4.9)) veroorzaakt een verschuiving van de wortel p = 16 met

$$\Delta p \approx 4.10^4.\ a_{15}\ \xi \approx 4.10^{14}\ \xi .$$

In het geval $\xi = \bar{\xi} = 2^{-40} \approx \frac{1}{2}10^{-12}$ (vgl. (4.10)) is dit ongeveer 200, een bedrag dat gezien de separatie van de wortels in de oorspronkelijke veelterm veel te groot is om de gepleegde eerste orde benadering in (39.6) te rechtvaardigen.

(39.8) Het blijkt echter wel duidelijk dat bij de polynomen met een dergelijke verdeling van nulpunten alleen al de afronding die nodig is om het polynoom in een rekenautomaat met een rekenprecisie van 12 decimalen te representeren de wortels onherkenbaar verminkt. De bepaling der eigenwaarden van een matrix met eigenwaarden in de buurt van 1, 2, 3, ... , 20 d.m.v. de oplossing van het expliciete karakteristieke polynoom is dus op een dergelijke machine niet mogelijk.

We wijzen er nog eens met nadruk op dat dit in laatste instantie niet veroorzaakt wordt door de afrondfouten tijdens de bepaling der coëfficiënten van $\Psi(A;\lambda)$ (die maken de zaak alleen nog maar

erger) maar door de onmogelijkheid de coëfficiënten adequaat in de machine te representeren.

(39.9) Gelukkig zijn er heel wat methoden die volgens andere principes te werk gaan.

§ 40 De machtsmethode.

(40.1) Zij A diagonaliseerbaar, $T^{-1}AT = D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$.
Neem aan dat $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$.
Zij $x_0 \in X$ een (in principe) willekeurige vector.

(40.2) X heeft een basis $v_1, \ldots, v_n$ van eigenvectoren van A (v_i eigenvector bij λ_i), dus we kunnen schrijven:

$$x_0 = a_1 v_1 + a_2 v_2 + \ldots + a_n v_n = a_1 v_1 + \tau_0.$$

(40.3) Wij bezien de rij geïtereerden $\{x_k\}_{k=0,1,\ldots}$ met

$$x_k = Ax_{k-1} \quad (k \geq 1).$$

Dan:

(40.4)
$$x_k = A^k x_0 = a_1 \lambda_1^k v_1 + \ldots\ldots + a_n \lambda_n^k v_n =$$
$$= \lambda_1^k (a_1 v_1 + \tau_k)$$

met

(40.5)
$$\tau_k = a_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k v_2 + \ldots\ldots + a_n \left(\frac{\lambda_n}{\lambda_1}\right)^k v_n.$$

(40.6) Omdat $\left|\frac{\lambda_i}{\lambda_1}\right| < 1$ voor $i \neq 1$ geldt $\lim \tau_k = 0$ en zal $\tau_k \to 0$ en dus zal x_k voor $k \to \infty$ steeds meer de richting van v_1, de "eerste" eigen vector aannemen (mits $a_1 \neq 0$).
Er geldt:

(40.7)
$$\lim \frac{\|x_{k+1}\|}{\|x_k\|} = |\lambda_1|$$

(40.8)
$$\lim \frac{(x_{k+1}, x_k)}{(x_k, x_k)} = \lambda_1$$

(40.9) In de praktijk zal men x_k door een geschikte factor delen om te voorkomen dat de coördinaten van x_k onbepaald groot of klein worden.
Het is gebruikelijk de geïtereerden te normeren op 1 (bv. in de 2- of ∞-norm).
Ook $y_k = \frac{x_k}{\|x_k\|}$ neemt steeds meer de richting van v_1 aan

en
$$\lim \frac{\|Ay_k\|}{\|y_k\|} = |\lambda_1|$$

$$\lim \frac{(Ay_k, y_k)}{(y_k, y_k)} = \lambda_1$$

(40.10) Uit (40.4) en (40.5) ziet men dat de convergentiesnelheid van het proces bepaald wordt door de verhouding $|\frac{\lambda_2}{\lambda_1}|$; indien $|\lambda_2| \approx |\lambda_1|$ kan de methode onder omstandigheden erg langzaam convergeren.
Voor manieren om de machtsmethode te versnellen zie bv. Wilkinson (Ch 9).

(40.11) Heeft men λ_1 en een bijbehorende eigenvector v_1 gevonden dan zou men met een startvector $x_0 \perp v_1$ in principe λ_2 (mits $|\lambda_2| > |\lambda_i|$ voor $i \geqslant 3$) kunnen vinden.
Iteratie met zo een x_o zal echter t.g.v. afrondfouten spoedig een vector opleveren die toch een component in de richting van v_1 heeft.

(40.12) Om toch de overige eigenwaarden (en eigenvectoren) van A te vinden zijn diverse deflatie-technieken ontwikkeld.
Daarbij wordt na bepaling van λ_1 en v_1 op speciale wijze uit A een matrix B geconstrueerd die eigenwaarden 0, $\lambda_2, \ldots, \lambda_n$ en dezelfde eigenvectoren $v_1, \ldots, v_n$ heeft. Etc..
Men moet wel bedenken dat steeds gedeflateerd wordt met benaderde eigenwaarden en -vectoren, hetgeen met name de ligging der nog te bepalen eigenvectoren aanmerkelijk kan beïnvloeden.(Het effect op de nog te bepalen eigenwaarden valt meestal wel mee).

(40.13) De grote hoeveelheid werk en de niet erg grote nauwkeurigheid zijn er de oorzaak van dat de machtsmethode thans niet veel meer gebruikt wordt, vooral omdat men nu over betere algorithmen beschikt.

(40.14) Een aan de machtsmethode verwant proces is dat der inverse- of Wielandt-iteratie.
Stel men heeft op een of andere wijze een benadering $\tilde{\lambda}_i$ van λ_i gevonden.
Als men aanneemt dat $\tilde{\lambda}_i$ dichter bij λ_i dan bij enige andere eigenwaarde ligt dan is $\lambda_i - \tilde{\lambda}_i$ de absoluut kleinste eigenwaarde van $A - \tilde{\lambda}_i I$ zodat $(\lambda_i - \tilde{\lambda}_i)^{-1}$ de absoluut grootste eigenwaarde van $(A - \tilde{\lambda}_i I)^{-1}$ is.

Men kan nu in principe de machtsmethode gaan toepassen op $(A - \tilde{\lambda}_i I)^{-1}$, maar om het onvoordelig inverteren te omzeilen (vgl(22.21)) gaat men als volgt te werk:

Kies een x_0
bepaal x_1 als oplossing van $(A - \tilde{\lambda}_i I)x_1 = x_0$
bepaal x_2 als oplossing van $(A - \tilde{\lambda}_i I)x_2 = x_1$
etc.

Het proces blijkt vaak bijzonder snel te convergeren en men kan aldus nog een verbetering van de oorspronkelijke $\tilde{\lambda}_i$ verkrijgen.

§ 41 Het gedrag van matrix - vector - iteratie.

(41.1) Informatie omtrent het gedrag van de rij vectoren $A^k x (k = 0,1,2,..)$ is van belang voor tal van (iteratieve) processen (zie bv. §28 §39 en §40).
Het is om deze reden dat wij er nader aandacht aan schenken.

(41.2) Zij
$$x_0 = x$$
$$x_k = A\, x_{k-1} \quad (k \geq 1)$$
Laat $B = T^{-1} AT$ de Jordannormaalvorm van A zijn (vgl. (19.16)), en definieer:
$$y_0 = T^{-1} x_0$$
(41.3)
$$y_k = B\, y_{k-1} = T^{-1}\, x_k \qquad k \geq 1$$
We mogen aannemen dat de diagonaalelementen $\lambda_1, \ldots, \lambda_n$ van B voldoen aan $|\lambda_1| \geq |\lambda_2| \geq \ldots .$

(41.4) Beschouw eerst de situatie dat B uit slechts één Jordan-kastje bestaat:
$$B = \begin{pmatrix} \lambda & 1 & & \Theta \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ \Theta & & & \lambda \end{pmatrix} \qquad (n \times n)$$
$$= \lambda\, I + N$$

(41.5) Toon aan dat $N^n = 0$ en dat
$$B^k = \begin{pmatrix} \lambda^k & \binom{k}{1}\lambda^{k-1} & - - - & \binom{k}{n-1}\lambda^{k-n+1} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \binom{k}{1}\lambda^{k-1} \\ \Theta & & & \lambda^k \end{pmatrix}$$

(41.6) Zij nu $y_0 = (\xi_1, \ldots, \xi_n)^T$ en zij l de hoogste index waarvoor $\lambda_l \neq 0$

Dan heeft $y_k = B^k y_0$ als eerste coördinaat:

$$\lambda^k \xi_1 + \binom{k}{1} \lambda^{k-1} \xi_2 + \ldots + \binom{k}{l-1} \lambda^{k-l+1} \xi_l$$

Voor $k \to \infty$ is alleen de laatste term hiervan belangrijk en men kan nog zeggen dat deze gelijk is aan

$$\frac{k^{l-1}}{(l-1)!} \lambda^{k-l+1} \xi_l \quad (1+0\left(\frac{1}{k}\right)).$$

Evenzo gedraagt zich voor $k \to \infty$ de tweede coördinaat van y_k als

$$\frac{k^{l-2}}{(l-2)!} \lambda^{k-l+2} \quad \xi_l(1+0\left(\frac{1}{k}\right))$$

en is dus klein t.o.v. de eerste coördinaat.

Idem voor de overige coördinaten.

Ga na dat men nu kan schrijven :

(41.7) $$y_k = k^{l-1}\lambda^k [I + 0\left(\frac{1}{k}\right)] \begin{pmatrix} c \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

waarin I de eenheidsmatrix is en $0(\frac{1}{k})$ een matrix voorstelt. Voorts

(41.8) $c = {}^{\xi_l}/\lambda^{l-1}.\ (l-1)!$

(41.9) Laat nu B uit meerdere Jordankastjes bestaan. Stel dat het i^e Jordankastje de afmeting $n_i \times n_i$ en als diagonaal element λ_i heeft.

Partitioneer y_0 op overeenkomstige wijze als B:

$$y_0 = (\xi_{11}, \ldots \xi_{1n_1}, \xi_{21} \ldots \xi_{2n_2}, \ldots.)^T.$$

Zij l_i de hoogste index waarvoor $\xi_{il_i} \neq 0$ is, en zij

$c_i = {}^{\xi_{l_i}}/\lambda_{n_i}^{l_i-1} (l_i-1)!$ (vgl. (41.8))

Dan geldt:

(41.10) $y_k = k^{l_1-1}\lambda_1^k[(I + O\left(\frac{1}{k}\right))]$

$$\begin{pmatrix} 1 & & \\ & k^{l_2-l_1}\left(\frac{\lambda_2}{\lambda_1}\right)^k & \\ & & k^{l_3-l_1}\left(\frac{\lambda_3}{\lambda_1}\right)^k \\ & & & \ddots \end{pmatrix} \begin{pmatrix} c_1 \\ 0 \\ \vdots \\ 0 \\ c_2 \\ 0 \\ \vdots \\ 0 \\ c_3 \\ 0 \\ \vdots \\ c \\ c \end{pmatrix}$$

(41.11) Stel nu dat de eerste j Jordankastjes allemaal dezelfde eigenwaarden hebben, dus $\lambda_1 = \lambda_2 = \ldots = \lambda_j$ en $|\lambda_j| > |\lambda_{j+1}|$ en laat c_1 t/m c_j niet allemaal 0 zijn. Dan spelen de overige vakjes praktisch niet meer mee. Er geldt:

(41.12) $y_k = k^{l_1-1}\lambda_1^k[I + O\left(\frac{1}{k}\right)](c_1, 0, \ldots, k^{l_2-l_1}c_2, 0\ldots0, \ldots, k^{l_j-l_1}c_j, 0 \ldots\ldots 0)^T$

$= \lambda_1 [I + O\left(\frac{1}{k}\right)] y_{k-1}$

(41.13) We bezien wat dit betekent voor de oorspronkelijke matrix A. Via (41.3) ziet men:

(41.14) $x_k = T y_k = \lambda_1 T[I + O\left(\frac{1}{k}\right)] T^{-1} x_{k-1} = \lambda_1[I + O\left(\frac{1}{k}\right)] x_{k-1}$.
en derhalve:

(41.15) $\|x_k\| \sim |\lambda_1| \|x_{k-1}\|$.

(41.16) Voor $k \to \infty$ verandert de "lengte" van een iterand bij elke stap met een factor die ongeveer gelijk is aan de spectraalstraal.

(41.17) <u>Opgave</u>. Ga na dat onder de gemaakte aannamen y_k en dus x_k voor $k \to \infty$ tot een zekere limietstand nadert.

(41.18) <u>Opgave</u>. Ga na dat (41.14) t/m (41.17) ook nog gelden als de eigenwaarden in eerste j Jordankastjes verschillend zijn maar wel dezelfde modulus hebben d.w.z. $|\lambda_1| = \ldots = |\lambda_j|$ en $|\lambda_j| > |\lambda_{j+1}|$, mits onder de $l_1, \ldots, l_j$ precies één grootste is.

(41.19) Opmerking. Tengevolge van afrondfouten in het iteratieve proces zullen de l_i vrijwel steeds hun maximale waarde hebben d.w.z. $l_i = n_i$.

(41.20) Opgave. Ga na dat voor elke matrix en elke beginvector x_0 geldt $\|x_k\| = |\lambda + o(1)|^k \|x_0\|$ voor zekere eigenwaarde λ (afh. van x_0).

§42 Methode van Jacobi

(42.1) Evenals de machtsmethode behoort de methode van Jacobi tot de derde categorie van §38.

(42.2) Zij $A = (a_{ij})$ een reële, symmetrische matrix.
A kan door een orthogonale transformatie op diagonaal vorm gebracht worden d.w.z. er is een orthogonale O zodat
$O^{-1} A O = D = \begin{pmatrix} \lambda_1 & & \Theta \\ & \ddots & \\ \Theta & & \lambda_n \end{pmatrix}$, met D een diagonaal matrix waarvan de diagonaalelementen juist de (reële) eigenwaarden van A zijn.

(42.3) De door <u>Jacobi</u> in 1846 aangegeven methode beoogt O successief op te bouwen als produkt van rotaties in geschikte vlakken.

(42.4) Er wordt gebruik gemaakt van de volgende eigenschap die we als opgave formuleren.

(42.5) <u>Opgave</u>. Voor een willekeurige matrix C en unitaire matrices U en V geldt $\|UCV\|_F = \|C\|_F$.
Bemerk hiertoe dat een kolom van UC dezelfde 2-norm heeft als de overeenkomstige kolom van C en dat een rij van CV dezelfde 2-norm heeft als de overeenkomstige rij van C.

(42.6) Zij nu voor $p < q$:

$$O_{pq} = \begin{pmatrix} 1 & & & & & & \Theta \\ & \ddots & & & & & \\ & & c & \cdots & s & & \\ & & \vdots & 1 \ddots 1 & \vdots & & \\ & & -s & \cdots & c & & \\ & & & & & \ddots & \\ \Theta & & & & & & 1 \end{pmatrix} \begin{matrix} \leftarrow p^e \\ \leftarrow q^e \end{matrix}$$

$\uparrow p^e$ (kolom) $\quad \uparrow q^e$ (kolom)

met $c^2 + s^2 = 1$

(42.7) Dan is O_{pq} orthogonaal en in feite een draaiing in het door de basisvectoren e_p en e_q opgespannen vlak.
Dus: $O_{pq}^{-1} = O_{pq}^T$.

(42.8) Zij a_{pq} een absoluut grootste buitendiagonaalelement van A.
Wegens symmetrie mogen we aannemen: $p < q$
Ontwerp nu:

(42.9)
$$B = O^T_{pq} \; A \; O_{pq}.$$

Deze matrix heeft natuurlijk dezelfde eigenwaarden als A.

(42.10) Opgave. Ga na dat $\sum\limits_{i,j=1}^{n} b^2_{ij} = \sum\limits_{i,j=1}^{n} a^2_{ij}$ (vgl. (42.5)).

(42.11) Opgave. Ga na dat B weer symmetrisch is en dat

$$\begin{aligned}
b_{ij} &= a_{ij} && \text{voor } i \neq p,\ j \neq q \\
b_{pj} &= c\, a_{pj} - s\, a_{qj} && \text{voor } j \neq p,q \\
b_{pp} &= c^2\, a_{pp} - 2cs\, a_{pq} + s^2\, a_{qq} && \\
b_{pq} &= cs\, (a_{pp} - a_{qq}) + (c^2 - s^2) a_{pq} && \\
b_{qj} &= s\, a_{pj} + c\, a_{qj} && \text{voor } j \neq p,q \\
b_{qp} &= b_{pq} && \\
b_{qq} &= s^2\, a_{pp} + 2cs\, a_{pq} + c^2\, a_{qq} &&
\end{aligned}$$

(42.12) Opgave. Ga na dat

$$\begin{pmatrix} c & -s \\ s & c \end{pmatrix} \begin{pmatrix} a_{pp} & a_{pq} \\ a_{qp} & a_{qq} \end{pmatrix} \begin{pmatrix} c & s \\ -s & c \end{pmatrix} = \begin{pmatrix} b_{pp} & b_{pq} \\ b_{qp} & b_{qq} \end{pmatrix}$$

(42.13) Opgave. Toon aan dat men c en s zo kan kiezen dat $b_{pq} = b_{qp} = 0$.

(42.14) Stel nu dat men c en s kiest als in (42.13). Dan geldt:

$$b^2_{pp} + b^2_{qq} = a^2_{pp} + 2\, a^2_{pq} + a^2_{qq}$$

(vgl. (42.5) en (42.12)).
Omdat $b_{ii} = a_{ii}$ voor $i \neq p,q$ geldt derhalve:

$$\sum_{i=1}^{n} b^2_{ii} = \sum_{i=1}^{n} a^2_{ii} + 2\, a^2_{pq}$$

(42.15) De kwadraatsom der diagonaalelementen is aldus met $2\, a^2_{pq}$ toegenomen.
De kwadraatsom der buitendiagonaalelementen moet dan met $2\, a^2_{pq}$ afgenomen zijn (vgl. (42.10)).

(42.16) Het proces wordt nu herhaald met B i.p.v. A.

Doorgaans zullen nulgemaakte elementen bij de volgende stap weer $\neq 0$ worden. De kwadraatsom der buitendiagonaalelementen blijft echter zakken.

(42.17) Men hoopt natuurlijk dat de kwadraatsom der buitendiagonaalelementen naar 0 gaat, waarna men eenvoudig van de diagonaal de eigenwaarden kan aflezen.
Dit zal zeker gebeuren als men telkens het absoluut grootste buitendiagonaalelement (vgl. (42.8)) nul maakt. Zij immers $N = n(n-1)$ het aantal buitendiagonaalelementen, en zij S_k de kwadraatsom der buitendiagonaalelementen na de k^e transformatiestap.

Dan geldt $\max\limits_{i \neq j} a_{ij}^2 \geq S_k/N$ zodat

$$S_{k+1} \leq S_k(1-\frac{2}{N}).$$

Dus $S_{k+1} \leq S_0(1-\frac{2}{N})^{k+1}$ en $S_{k+1} \to 0$ voor $k \to \infty$.

(42.18) Dit laatste zou overigens aanleiding kunnen zijn tot sombere gedachten over de convergentiesnelheid, maar dat valt mee. Men kan zelfs aantonen (moeilijk!) dat Jacobi superlineair convergeert.

(42.19) In de praktijk blijken na circa $3n^2$ transformatiestappen de buitendiagonaalelementen bij een rekenprecisie van 12 decimalen doorgaans niet meer significant van 0 te verschillen (empirisch). Aangezien elke stap circa $4n$ AV. kost is de totale hoeveelheid werk dus omstreeks $12n^3$ AV . Deze voorstelling van de hoeveelheid rekenwerk is echter zeer bedriegelijk. Voor elke transformatiestap moet men immers eerst het abs. grootste buitendiagonaalelement gaan bepalen. Bij het met de hand rekenen ziet men dat grootste element op het oog, de rekenautomaat kan het grootste van $\frac{1}{2}n(n-1)$ elementen echter alleen d.m.v. $\frac{1}{2}n(n-1)-1$ aftrekkingen bepalen, dat is veel meer werk dan de $4n$ AV . die vervolgens het uitvoeren van de transformatiestap kost! Vandaar dat men maar niet zoekt, maar gewoon achtereenvolgens a_{12}, a_{13}, ..., a_{23},...,a_{3n},...,$a_{n-1,n}$ nul maakt en dan weer van voren af aan begint (seriëel Jacobi proces). Merk op dat het nu helemaal niet meer duidelijk is dat het proces nog convergeert.
Men kan het echter wel bewijzen (niet eenvoudig).

(42.20) Corbato merkte op dat men om bij het oorspronkelijke Jacobi-proces het abs. grootste buitendiagonaalelement te bepalen toch niet bij elke stap $\frac{1}{2}n(n-1)-1$ aftrekkingen hoeft uit te voeren.
Per stap veranderen immers slechts 2 rijen en kolommen van de matrix.
Door bij elke rij de plaats en de grootte van het abs. grootste element van die rij te onthouden, heeft men per transformatiestap slechts circa 2n aftrekkingen nodig om het abs. grootste buitendiagonaalelement te bepalen. Het is niet bekend of op deze wijze het echte Jacobiproces toch niet sneller convergeert dan het seriële proces.

(42.21) De numerieke stabiliteit van het Jacobiproces is goed.
Stel dat men bij het rekenen in eindige precisie (dus na ongeveer $3n^2$ stappen voor het geval van 12 decimalen) uitkomt op

$$\begin{pmatrix} \lambda_1^* & & \sim\Theta \\ & \ddots & \\ \sim\Theta & & \lambda_n^* \end{pmatrix}$$

terwijl men bij exact rekenen had moeten krijgen:

$$\begin{pmatrix} \lambda_1 & & \Theta \\ & \ddots & \\ \Theta & & \lambda_n \end{pmatrix}$$

Dan kan men aantonen dat geldt:

$$\left(\frac{\sum_{i=1}^{n} (\lambda_i^*-\lambda_i)^2}{\Sigma\lambda_i^2}\right)^{\frac{1}{2}} < 108n^{\frac{3}{2}}\ \bar{\xi}$$

Voor een 20 x 20 matrix betekent dit een verlies van 4 decimale cijfers. Dit is echter een absolute bovengrens, bij de bepaling waarvan men er rekening mee gehouden heeft dat bij elke arithmetische operatie de maximale fout wordt gemaakt. Doordat de werkelijke optredende fouten doorgaans (ondanks hun gedetermineerdheid, ze zijn elke keer dat men hetzelfde proces draait, hetzelfde) een stochastisch karakter lijken te hebben, treedt een soort statistisch effect op en wordt de werkelijke fout beter geschat door $11\ n^{\frac{3}{4}}\ \bar{\xi}$.

(42.22) Het Jacobiproces is niet het snelst bekende proces, wel het eenvoudigste, en daardoor voor niet te grote matrices zeer aantrekkelijk.

(42.23) Bovendien is het aantrekkelijk dat men meteen de eigenvectoren verkrijgt, nl. als de kolommen van het produkt der O_{pq} (in de goede volgorde).

(42.24) Merk nog op dat men m.b.v. (36.12) kan zien hoe het ≠ 0 zijn der buitendiagonaalelementen de eigenwaarden beïnvloedt en m.b.v. (37.28) hoe het met de eigenvectoren staat.

(42.25) Opmerking. In principe kan men Jacobi toepassen op willekeurige normale matrices (vgl. (19.10)).
Een recent algorithme van Paardekooper transformeert stapsgewijs een willekeurige (niet noodzakelijk normale) matrix tot een matrix die bijna normaal is waarna men vervolgens Jacobi toepast om een (goede) benadering der eigenwaarden te vinden.
M.H.C. Paardekooper - An Eigenvalue Algorithm based on Norm - Reducing Transformations. Diss Eindhoven 1969.

(42.26) Litteratuur

Wilkinson, J.H. - The Algebr. Eigenvalue Probl. (265-282).
Corbato, F.J. - JACM 10(1963), 123-125.
van Kempen, H.P.M. - Num. Math. 9(1966), 11-22.

§ 43 Methode van Householder - Givens.

(43.1) De methode van Householder - Givens behoort tot de tweede categorie van §38.
Zij construeert in eindig veel stappen een orthogonale matrix O die de reële symmetrische matrix A naar tridiagonaalvorm transformeert.
De methode is oorspronkelijk ontwikkeld door Givens; later heeft Householder een andere (snellere) manier aangegeven om de reductie tot tridiagonaalvorm tot stand te brengen.

(43.2) Bij Givens wordt O stapsgewijs opgebouwd uit draaiingen zoals in (42.6). Iedere draaiing is bedoeld om 2 symmetrisch gelegen matrixelementen buiten de tridiagonale band nul te maken, waarbij elementen die eenmaal nulgemaakt zijn nul blijven gedurende het verdere verloop van het proces.

De methode gaat als volgt:
-eerst worden a_{31} en a_{13} nulgemaakt door een rotatie in het (e_2, e_3)-vlak. Resulterende matrix $B = (b_{ij})$
-dan worden b_{41} en b_{14} nulgemaakt door een rotatie in het (e_2, e_4)-vlak.
-etc.

(43.3) Opgave. Ga na dat nulgemaakte elementen inderdaad nul blijven.

(43.4) Bij Householder wordt O stapsgewijs opgebouwd uit Householder-transformaties $H = I - 2w.w^T$ waarin w een vector met 2-norm 1 is (vgl. (33.3)).

In de eerste stap Householder bepaalt men H_1 (dus eigenlijk w) zó dat $a_{31}, a_{41}, \ldots, a_{n1}$ en $a_{13}, a_{14}, \ldots, a_{1n}$ alle tegelijk nul worden. Resulterende matrix: $B = (b_{ij})$.
In de tweede stap Householder bepaalt men H_2 zo dat de elementen $b_{42}, \ldots, b_{n2}$ en $b_{24}, \ldots, b_{2n}$ nul worden (en de elementen $b_{31}, \ldots, b_{n1}$ en $b_{13}, \ldots, b_{1n}$ nul blijven). Etc.

(43.5) De benodigde rekentijd voor de reductie tot tridiagonaalvorm is beperkt : $\frac{4}{3}n^3$ AV bij Givens en $\frac{2}{3}n^3$ AV bij Householder.

(43.6) Beide algorithmen zijn zeer stabiel.
Noteren we met λ_i resp. λ_i^* de (naar grootte gerangschikte) eigenwaarden van de oorspronkelijke en van de in eindige precisie berekende tridiagonaalmatrix dan geldt

$$\left(\frac{\sum_{i=1}^{n} (\lambda_i^* - \lambda_i)^2}{\sum \lambda_i^2} \right)^{\frac{1}{2}} \leq 17\, n^{3/2}\, \xi \qquad \text{bij Givens}$$

en

$$\text{"} \qquad \leq 6n^2\, \xi \qquad \text{bij Householder}$$

(vergelijk ook (42.21)).

(43.7) Zij T de verkregen tridiagonaalmatrix en stel

$$T = \begin{pmatrix} a_1 & b_1 & 0 & \cdots & & 0 \\ b_1 & a_2 & b_2 & & & \vdots \\ 0 & b_2 & a_3 & \ddots & & \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ & & & \ddots & \ddots & b_{n-1} \\ 0 & \cdots & & 0 & b_{n-1} & a_n \end{pmatrix}$$

De waarde van $\Psi(T;\lambda) = \det(T-\lambda I)$ is nu heel goedkoop te bepalen via een recurrente betrekking.

(43.8) Definiëer nl.

$$p_i(\lambda) = \begin{vmatrix} a_1-\lambda & b_1 & & \\ b_1 & a_2-\lambda & \ddots & \\ & \ddots & \ddots & b_{i-1} \\ & & b_{i-1} & a_i-\lambda \end{vmatrix} \qquad (i \geq 1)$$

Met $p_{-1}(\lambda) = 0$, $p_0(\lambda) = 1$ krijgt men:

(43.9) $$p_i(\lambda) = (a_i - \lambda)\, p_{i-1}(\lambda) - b_{i-1}^2\, p_{i-2}(\lambda) \qquad (i \geq 1)$$

(43.10) Opgave. Verifiëer (43.9) en ga na dat $\Psi(T;\lambda) = p_n(\lambda)$.

(43.11) Evenwel, in plaats van nu bijv. Koorden - Newton toe te passen om de nulpunten van $\Psi(T;\lambda)$ te bepalen, gebruikt men de rij $\{p_i(\lambda)\}$ op een andere manier.
Onder aanname dat alle $b_i \neq 0$(wat betekent $b_i = 0$?) vormt de rij $p_0, p_1, \ldots, p_n$ een zgn. Sturm-rij van polynomen, d.w.z. geen van de p_i is $\equiv 0$, en in een nulpunt θ van $p_i (i \geq 1)$ geldt $\operatorname{sgn}(p_{i-1}(\theta)) = -\operatorname{sgn}(p_{i+1}(\theta)) \neq 0$, Zie verder §44

(43.12) Opgave. Verifiëer dit laatste m.b.v. (43.9)

(43.13) Met behulp van de Sturm-rij kan men een heel bruikbaar procédé voor de bepaling van de eigenwaarden van T ontwerpen.
Verdere bijzonderheden vindt men bv. bij Wilkinson, Ortega en Givens.

(43.14) Het bepalen van de eigenvectoren van een tridiagonaalmatrix is een vervelende zaak. Hiervoor is nl. geen stabiele algorithme bekend. (vgl. Wilkinson p. 315 e.v.). Soms gebruikt men Wielandt iteratie.

§44 Tridiagonaalmatrices en Sturm-rijen

$$C_n = \begin{pmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \beta_3 & & \\ & \ddots & \ddots & \ddots & \\ & & & & \beta_n \\ & & & \beta_n & \alpha_n \end{pmatrix}$$

(44.1) Als alle $\beta_i \neq 0$ en det $C_n = 0$ dan is de vgl. $C_n x = 0$ eenduidig oplosbaar op een scalaire factor na.
Hint: onderscheid $x_1 = 0$ en $x_1 \neq 0$

(44.2) Als wel sommige $\beta_i = 0$ (zeg k der β_i zijn 0) dan is C_n directe som van $k+1$ tridiag matrices in elk waarvan alle β_i ongelijk 0. Dan is de nulruimte van C_n directe som van de nulruimten van elk der afzonderlijke matrices, dus hoogstens $k+1$ dimensionaal.
Gevolg:

(44.3) Als C_n een μ-voudige eigenwaarde heeft moeten minstens $\mu-1$ der β_i nul zijn. Dus alle $\beta_i \neq 0 \Rightarrow$ alle eigenwaarden enkelvoudig (niet andersom!).

(44.4) Zij $p_i(\lambda) = \det(C_i - \lambda)$. Dan geldt met $p_0(\lambda) = 1$:

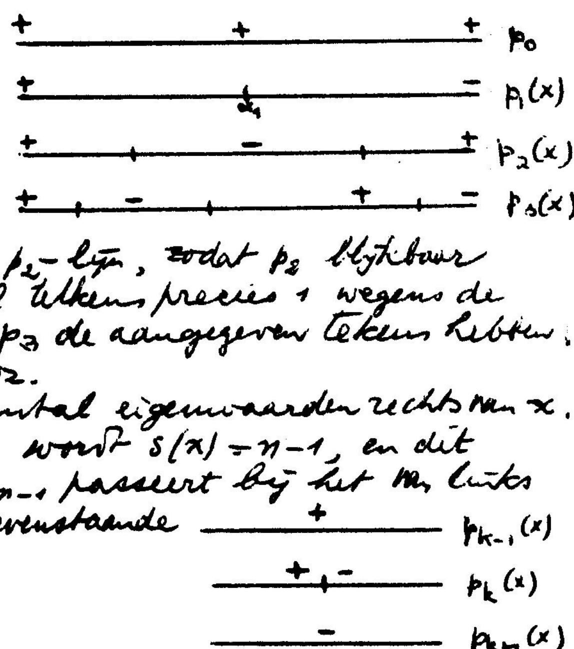
$$p_1(\lambda) = \alpha_1 - \lambda$$
$$p_2(\lambda) = (\alpha_2 - \lambda) p_1(\lambda) - \beta_2^2 p_0(\lambda) \quad \text{en algemener}$$
$$p_i(\lambda) = (\alpha_i - \lambda) p_{i-1}(\lambda) - \beta_i^2 p_{i-2}(\lambda)$$

(zo kan men dus $\det(C_n - \lambda)$ gemakkelijk uitrekenen)

Laat nu alle $\beta_i \neq 0$ (dit is geen beperking; ga na)

(44.5) Stelling (Sturm): Zij $s(x)$ het aantal overeenstemmingen in teken van de opeenvolgende termen van de rij $p_0(x), p_1(x), \dots p_n(x)$. Als $p_k(x) = 0$ kennen we hem het tegengestelde teken toe van $p_{k-1}(x)$ (die dan $\neq 0$) Dan is $s(x)$ het aantal eigenwaarden dat strikt groter dan x is.

Bewijs. De rij $p_i(x)$ is geheel positief voor x sterk negatief, en alternerend voor x sterk positief. Dit is in de tekening aangegeven. Voorts: als $p_k(x) = 0$ dan zijn $p_{k-1}(x)$ en $p_{k+1}(x) \neq 0$ en verschillend van teken.



Wegens $p_1(\alpha_1) = 0$ ontstaat het − teken op de p_2-lijn, zodat p_2 blijkbaar nulpunten heeft rechts en links van α_1, en wel telkens precies 1 wegens de graad van p_2. Bij deze nulpunten moet dan p_3 de aangegeven tekens hebben, waaruit dan weer nulpunten volgen enz.

Voor x sterk negatief is $s(x) = n$ = aantal eigenwaarden rechts van x. Als x = meest linkse nulpunt van $p_n(x)$ wordt $s(x) = n-1$, en dit blijft zo als x een nulpunt van p_0 t/m p_{n-1} passeert bij het van links naar rechts lopen (ga dit aan de hand van nevenstaande prent na. Dit voltooit het bewijs

§45. Eigenwaarden van niet-normale matrices

Het eigenprobleem voor niet-normale matrices is veel moeilijker dan dat voor normale. Een der beste thans beschikbare methodes is de QR methode. Alvorens deze te beschrijven beschouwen we eerst een generalisatie van de machtsmethode (matrix-vector iteratie).

We nemen eerst aan dat de eigenwaarden van de $n \times n$ matrix A allemaal verschillende modulus hebben: $|\lambda_1|>|\lambda_2|>|\lambda_3|>\dots>|\lambda_n|$ met bijbehorende eigenvectoren $g_1, g_2, \dots, g_n$. Zij G de matrix $(g_1, g_2, \dots, g_n)$ en zij $G = UR$, U unitair, R bovendriehoeks, d.w.z. U wordt verkregen door G van links naar rechts te orthonormaliseren. De k-de kolom van U wordt met u_k aangeduid.

Zij P een matrix waarvan de kolommen lineair onafhankelijk zijn. Zij $sp(P)_{1,k}$ de ruimte opgespannen door de eerste k kolommen van de matrix P. We noemen P consistent met G wanneer $sp(P)_{1,k} \cap sp(G)_{k+1,n} = \{0\}$ voor alle k. M.a.w. als men een vector uit $sp(P)_{1,k}$ schrijft als lineaire combinatie van $g_1, \dots, g_n$ dan mogen de coordinaten t.o.v. $g_1, \dots, g_k$ niet alle 0 zijn tenzij het de nulvector betreft.
Blijkbaar is P consistent met G d.d.d. als $P = GC$ waarin C een matrix is waarvan alle leidende hoofdminoren $\neq 0$.
Als P consistent is met G dan is ook PB consistent met G wanneer B een bovendriehoeksmatrix is waarvan alle diagonaalelementen ongelijk 0 zijn, immers $sp(PB)_{1,k} = sp(P)_{1,k}$ voor alle k.

Zij nu X_0 een matrix consistent met G en beschouw de rij

(1) $$X_{i+1} = AX_i$$

Merk op dat als $\lambda_n \neq 0$ dan alle X_i consistent met G zijn, en dat als $\lambda_n = 0$ de eerste $n-1$ kolommen van alle X_i consistent met de eerste $n-1$ kolommen van G zijn. Immers als $X_0 = GC$ dan is $A^i X_0 = G D^i C$, waarin D de diagonaalmatrix der eigenwaarden is.

We beschrijven nu eerst iets heuristisch wat er gebeurt. Zij $(X_i)_k$ de k-de kolom van X_i. Dan geldt

(2) $$\begin{aligned} (X_i)_1 &= a_1 \lambda_1^i g_1 + a_2 \lambda_2^i g_2 + \dots \\ (X_i)_2 &= b_1 \lambda_1^i g_1 + b_2 \lambda_2^i g_2 + \dots \\ (X_i)_3 &= c_1 \lambda_1^i g_1 + c_2 \lambda_2^i g_2 + \dots \end{aligned}$$

met $a_1 \neq 0$, $\det\begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \end{pmatrix} \neq 0$, $\det\begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{pmatrix} \neq 0$ etc. wegens de consistentie.

Bijgevolg nadert $sp(X_i)_1$ naar $sp(G)_1$ (machtsmethode), dus naar $sp(U)_1$.
$sp(X_i)_{1,2}$ heeft als basis $g_1 + o(1)g_2 + o(1)g_3 + \dots$ en $g_2 + o(1)g_3 + \dots$
en dus nadert $sp(X_i)_{1,2}$ naar $sp(G)_{1,2} = sp(U)_{1,2}$.
$sp(X_i)_{1,3}$ nadert evenzo tot $sp(G)_{1,3} = sp(U)_{1,3}$.

Derhalve, als $(W_i)_1$ een eenheidsvector is met richting $(X_i)_1$ dan nadert deze op een scalaire factor met modulus 1 na tot u_1.
Als $(W_i)_2$ een eenheidsvector is in $sp(X)_{1,2}$ orthogonaal op $(X_i)_1$, dan nadert deze op een scalaire factor met modulus 1 na tot u_2. Etc

Conclusie: zij W_i een van links naar rechts georthonormaliseerde van X_i, dan geldt $W_i \approx U E_i$ voor een rij diagonaal matrices E_i met diagonaalelementen van modulus 1, zodat $W_i^H A W_i \approx E_i^H U^H A U E_i = E_i^H R G^{-1} A G R^{-1} E_i$ en dit nadert naar een bovendriehoeksmatrix waarin op de diagonaal de eigenwaarden in volgorde van dalende modulus zijn af te lezen. Merk echter op dat $W_i^H A W_i$ niet echt een limiet hoeft te hebben: met toenemende i gaan slechts de moduli der elementen naar een limiet.

We bekijken de zaak nu preciezer. Wanneer we een matrix X van links naar rechts moeten orthonormaliseren krijgt men op rechtsvermenigvuldiging met een unitaire diagonaal matrix na (zoals E_i) hetzelfde resultaat als wanneer men XB van links naar rechts orthonormaliseert, B als boven.

Dus in plaats van X_i te orthonormaliseren mogen we ook wel Y_i orthonormaliseren met Y_i als onder (zie (2))

(3)
$$(Y_i)_1 = g_1 + O\left(\frac{\lambda_2}{\lambda_1}\right)^i g_2 + O\left(\frac{\lambda_3}{\lambda_1}\right)^i g_3 + \cdots$$
$$(Y_i)_2 = g_2 + O\left(\frac{\lambda_3}{\lambda_2}\right)^i g_3 + \cdots$$
$$(Y_i)_3 = g_3 + O\left(\frac{\lambda_4}{\lambda_3}\right)^i g_4 + \cdots$$

Hierbij is weer gebruikt dat in elk geval de eerste $n-1$ kolommen van X_i consistent zijn met de eerste $n-1$ kolommen van G. Aangezien de overgang naar (3) neerkomt op rechtsvermenigvuldiging met een bovendriehoeksmatrix met niet verdwijnende diagonaalelementen (zie B hierboven) zijn ook de eerste $n-1$ kolommen van de aldus verkregen matrix Y consistent met die van G. Deze opmerkingen gelden ook voor de nog volgende herleidingen.

Aangezien g_k een lin. combinatie is van $u_1, \ldots u_k$ met de coefficient van u_k ongelijk 0 geldt

(4)
$$(Y_i)_1 = (a_1' + o(1))u_1 + O\left(\frac{\lambda_2}{\lambda_1}\right)^i u_2 + O\left(\frac{\lambda_3}{\lambda_1}\right)^i u_3 + \cdots \qquad a_1' \neq 0$$
$$(Y_i)_2 = (b_1' + o(1))u_1 + (b_2' + o(1))u_2 + O\left(\frac{\lambda_3}{\lambda_2}\right)^i u_3 + O\left(\frac{\lambda_4}{\lambda_2}\right)^i u_4 + \cdots \qquad b_2' \neq 0$$
$$(Y_i)_3 = (c_1' + o(1))u_1 + (c_2' + o(1))u_2 + (c_3' + o(1))u_3 + O\left(\frac{\lambda_4}{\lambda_3}\right)^i u_4 + \cdots \qquad c_3' \neq 0.$$

Door nu $(Y_i)_2$ te verminderen met een geschikt veelvoud van $(Y_i)_1$, evenzo $(Y_i)_3$ met een geschikt veelvoud van $(Y_i)_1$ en $(Y_i)_2$, en vervolgens nog te delen door een geschikte factor (deze operaties komen juist neer op rechtsvermenigvuldiging met een matrix B als boven) mogen we dus ook wel orthonormaliseren

(5)
$$(Z_i)_1 = u_1 + O\left(\frac{\lambda_2}{\lambda_1}\right)^i u_2 + O\left(\frac{\lambda_3}{\lambda_1}\right)^i u_3 + \cdots$$
$$(Z_i)_2 = u_2 + O\left(\frac{\lambda_3}{\lambda_2}\right)^i u_3 + \cdots$$
$$(Z_i)_3 = u_3 + O\left(\frac{\lambda_4}{\lambda_3}\right)^i u_4 + \cdots$$

Wegens $\|(Z_i)_1\| = 1 + o(1)$ in de 2-norm is $((Z_i)_2, (Z_i)_1)/\|(Z_i)_1\|^2 = O\left(\frac{\lambda_2}{\lambda_1}\right)^i$ zodat het ongenormaliseerde orthogonalisatie resultaat luidt

$(V_i)_1 = (Z_i)_1$

(6) $(V_i)_2 = O(\frac{\lambda_2}{\lambda_1})^i u_1 + (1+o(1))\, u_2 + O(\frac{\lambda_3}{\lambda_2})^i u_3 + \cdots$ en $\|(V_i)_2\| = 1+o(1)$

$(V_i)_3 = (Z_i)_3 - [((Z_i)_3,(V_i)_1)/\|(V_i)_1\|^2](V_i)_1 - [((Z_i)_3,(V_i)_2)/\|(V_i)_2\|^2](V_i)_2 =$

$= O(\frac{\lambda_3}{\lambda_1})^i u_1 + O(\frac{\lambda_3}{\lambda_2})^i u_2 + (1+o(1))\, u_3 + O(\frac{\lambda_4}{\lambda_3})^i u_4 + \cdots$ en $\|(V_i)_3\| = 1+o(1)$

Wegens de aangegeven normen der $\|(V_i)_k\|$ hebben de uiteindelijk verkregen orthonormale vectoren $(W_i)_k$ dezelfde gedaante als $(V_i)_k$ in (6) op een scalaire factor met modulus 1 na.

Hieruit volgt $W_i = U T_i E_i$, met E_i een unitaire diagonaalmatrix en T_i als onder:

$$(7)\qquad T_i = \begin{pmatrix} 1+o(1) & O(\frac{\lambda_2}{\lambda_1})^i & O(\frac{\lambda_3}{\lambda_1})^i & O(\frac{\lambda_4}{\lambda_1})^i \\ O(\frac{\lambda_2}{\lambda_1})^i & 1+o(1) & O(\frac{\lambda_3}{\lambda_2})^i & O(\frac{\lambda_4}{\lambda_2})^i \\ O(\frac{\lambda_3}{\lambda_1})^i & O(\frac{\lambda_3}{\lambda_2})^i & 1+o(1) & O(\frac{\lambda_4}{\lambda_3})^i \\ O(\frac{\lambda_4}{\lambda_1})^i & O(\frac{\lambda_4}{\lambda_2})^i & O(\frac{\lambda_4}{\lambda_3})^i & 1+o(1) \end{pmatrix}$$

Hieruit volgt $W_i^H A W_i = E_i^H T_i^H U^H A U T_i E_i = E_i^H T_i^H R G^{-1} A G R^{-1} T_i E_i = E_i^H T_i^H S T_i E_i$ met S een bovendriehoeksmatrix met op de diagonaal de eigenwaarden van A in volgorde van dalende modulus. Men vindt

$$(8)\qquad W_i^H A W_i = E_i^H \begin{pmatrix} * & * & * & * \\ O(\frac{\lambda_2}{\lambda_1})^i & * & * & * \\ O(\frac{\lambda_3}{\lambda_1})^i & O(\frac{\lambda_3}{\lambda_2})^i & * & * \\ O(\frac{\lambda_4}{\lambda_1})^i & O(\frac{\lambda_4}{\lambda_2})^i & O(\frac{\lambda_4}{\lambda_3})^i & * \end{pmatrix} E_i$$

In feite is de afleiding vanaf (6) slechts goed als $\lambda_n \neq 0$. Het resultaat is echter ook juist voor $\lambda_n = 0$. Immers dan is $\lambda_{n-1} \neq 0$ en zijn (3) t/m (6) in orde t/m index $n-1$. $(Y_i)_n$, $(Z_i)_n$ en $(V_i)_n$ worden nu echter 0. Evenwel zijn $(Y_i)_k$, $(Z_i)_k$ en $(V_i)_k$, $k = 1, \ldots, n-1$ nu onafhankelijke lineaire combinaties van $g_1, \ldots, g_{n-1}$, dus van $u_1, \ldots, u_{n-1}$, zodat we wel verplicht zijn $(W_i)_n = u_n$ te nemen op een scalaire factor met modulus 1 na. Daardoor ontstaan in (7) rechtsonder een 1 en overigens nullen in de laatste rij en kolom. (7) en (8) blijven dus goed.

<u>Practische uitvoering</u>

Het is duidelijk niet zinvol na vele slagen X_i eens te orthogonaliseren omdat dan inmiddels alle kolommen zowat de richting van g_1 hebben gekregen, en dus de overige eigenvectoren nauwelijks meer vertegenwoordigd zijn.

Men mag echter ook wel na elke stap orthonormaliseren. Orthonormaliseren van een matrix X betekent immers het bepalen van een unitaire W en een (eventueel singuliere) bovendriehoeksmatrix R zodat $WR = X$. Als dus in elke stap aldus geldt $\tilde{W}_{i+1} R_{i+1} = A \tilde{W}_i$, dan geldt ook $\tilde{W}_{i+1} R_{i+1} R_i \cdots R_1 = A^i W_0$, zodat $\tilde{W}_{i+1}$ nog steeds een georthonormaliseerde van $A^i W_0$ is, en dus $\tilde{W}_{i+1} = W_{i+1}$ op een unitaire diagonaalmatrix na, wegens de reeds eerder gebleken eenduidigheid van het orthogonaliseringsproces in dit geval.

(NB De toevoeging "in dit geval" slaat hierop dat in $A^i X_0$ de eerste $n-1$ kolommen zeker onafhankelijk zijn; als dit eens niet zo was, bijv. als we eens de 0-matrix zouden orthogonaliseren dan kunnen we met de hier gebruikte definitie M_i orthogonaliseren een willekeurige unitaire matrix krijgen, met uiteraard een matrix R die 0 is)

Dus door na elke stap te orthogonaliseren krijgen we in feite ook de rij W_i en kunnen dan $W_i^H A W_i$ uitrekenen (eventueel af en toe, want dit is natuurlijk nogal tijdrovend).

De matrix $M_i \underset{D}{=} W_i^H A W_i$ is van zo groot belang dat we hier graag rechtstreeks relaties voor willen hebben, en hem niet telkens apart willen uitrekenen. Wegens $AW_i = W_{i+1} R_{i+1}$ geldt $M_i = W_i^H W_{i+1} R_{i+1}$, dwz. dat R_{i+1} verkregen wordt ~~bij het~~ van links naar rechts orthonormaliseren van M_i. Met $Q_{i+1} = W_i^H W_{i+1}$ geldt dan $M_i = Q_{i+1} R_{i+1}$. En $M_{i+1} = W_{i+1}^H A W_{i+1} = Q_{i+1}^H W_i^H A W_i Q_{i+1} = Q_{i+1}^H M_i Q_{i+1} = R_{i+1} Q_{i+1}$.

Zo krijgen we de zeer bekende en zeer veel gebruikte

<u>QR-algoritme</u> Zij W_0 een unitaire matrix die consistent is met G. Zij $M_0 = W_0^H A W_0$. Genereer nu een rij matrices M_i als volgt: orthonormaliseer M_i, dwz. bepaal Q_{i+1} (unitair) en R_{i+1} (bovendriehoeks) zodat $M_i = Q_{i+1} R_{i+1}$ en bereken M_{i+1} als $R_{i+1} Q_{i+1}$. Dan heeft M_i de gedaante van de matrix in (8) zodat dus de rij matrices M_i naar een bovendriehoekgedaante gaat met de eigenwaarden op de diagonaal.

Er blijven nog ettelijke vragen te beantwoorden:
a) hoe bedenk je een W_0 die consistent is met G
b) hoe staat het met de hoeveelheid werk, aangezien de convergentiesnelheid afhangt van de verhouding der eigenwaarden
c) wat gebeurt er als meerdere eigenwaarden dezelfde modulus hebben.

Ad. a <u>Startmatrix</u> W_0. Dit probleem wordt zeer bevredigend opgelost door:

<u>Stelling</u>. Bepaal W_0 zo dat $M_0 = W_0^H A W_0$ de Hessenberg gedaante heeft (dwz. dat de elementen op de plaatsen (i,j) met $i > j+1$ nul zijn). Als M_0 dan geen der elementen $(i, i-1)$ nul heeft dan is W_0 consistent met G.

<u>Opmerking 1</u>: Als wel een of meer elementen op plaatsen $(i, i-1)$ nul zijn reduceert het probleem tot 2 of meer kleinere eigenwaarde problemen waarvoor dan wel alle elementen op plaatsen $(i, i-1)$ ongelijk 0 zijn.
<u>Opmerking 2</u> De Hessenberg gedaante is eenvoudig te bereiken, namelijk met Householder transformaties op dezelfde wijze als waarop een symmetrische matrix op de tridiagonaalgedaante gebracht wordt.

Bewijs van de stelling. Stel dat W_0 niet consistent is met G. Zij dan k het kleinste natuurlijke getal zodat $sp(W_0)_{1,k} \cap sp(G)_{k+1,n} \neq \{0\}$. Dan is er dus een vector x met k-de coordinaat $\neq 0$ en $(k+1)$-ste t/m n-de coordinaat 0 zodat $W_0 x \in Z = sp(G)_{k+1,n}$. Dan ook $A^i W_0 x \in Z$ dus ook $W_0 M_0^i x (= A^i W_0 x) \in Z$. Nu heeft $M_0^i x$ de $(k+i)$-de coordinaat $\neq 0$ en alle volgende coordinaten 0. Dus $x, M_0 x, \ldots, M_0^{n-k} x$ zijn lin. onafh. Dus ook $W_0 x, W_0 M x, \ldots W_0 M^{n-k} x$ zijn lin. onafh, het zijn er $n-k+1$, ze liggen in Z, die $n-k$ dimensionaal is, en dat is een tegenspraak.

Ad b. Hoeveelheid werk. Werken met de Hessenberg gedaante is ook bijzonder gunstig voor de hoeveelheid werk. De QR ontbinding van M_0 voert men namelijk uit door M_0 van links met vlakke draaiingen in resp de vlakken $(1,2), (2,3), \ldots, (n-1,n)$ te vermenigvuldigen waardoor successievelijk de codiagonaalelementen 0 gemaakt worden. Dit kost $2n^2$ operaties. De RQ compositie van M_1 wordt dan tot stand gebracht door R van rechts met de inversen der draaiingen te vermenigvuldigen, weer ten koste van $2n^2$ operaties. Dus $4n^2$ per complete QR slag. Het merkwaardige is dat de Q dus niet ontstaat. Voorts heeft M_1 weer Hessenberg gedaante, en zo gaan we door.

De numerieke stabiliteit is aldus uitstekend omdat wegens $M_{i+1} = Q_{i+1}^H M_i Q_i$ en Q_i een product van vlakke draaiingen Wilkinson's stabiliteitstheorie van toepassing is (Wilkinson p140) die zegt dat in dit geval de verkregen matrix dezelfde eigenwaarden heeft als een matrix die vlak bij de beginmatrix ligt

Versnelling. Uit (8) zien we dat rechtsonderaan de absoluut kleinste eigenwaarde verschijnt. We kunnen het proces nu trachten te versnellen door in M_i alle diagonaal elementen te verminderen met het rechtsonder element; als dit laatste een goede benadering is voor λ_n zal de λ_n van de nieuwe matrix veel kleiner zijn en dus λ_n/λ_1 t/m λ_n/λ_{n-1} ook veel kleiner, waardoor de convergentie veel sneller wordt. Over dit soort versnellingstechnieken waarmee men wel kubische convergentie nastreeft is een hele literatuur ontstaan.

Ad c. Meervoudige eigenwaarden. Zij $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$ en beschouw de eigenwaarden ~~van gelijke~~ modulus als één groep. Partitioneer A, G en alle overige matrices volgens deze indeling in groepen. Dan gaat de afleiding van (8) door waarin (8) weer als gepartitioneerde matrix moet worden beschouwd.

Men werkt weer met de Hessenberg gedaante, waarbij nu niet meer alle codiagonaal elementen naar 0 hoeven te gaan, maar alleen die codiagonaalelementen die op de partitiepunten liggen.

Voor de versnelling pleegt men nu naar de 2×2-matrix rechtsonder te kijken, en gebruikt de eigenwaarden daarvan als benadering van de 2 eigenwaarden van kleinste modulus om weer alle eigenwaarden te verschuiven. Er zijn slimme methodes bedacht om bij reële matrices met complexe wortels de verschuiving toch zo uit te voeren dat de matrix reëel blijft, hetgeen prettig is omdat computers met reële getallen veel sneller rekenen dan met complexe.

Convergentie pleegt zeer bevredigend te zijn. Algemene bewijzen ontbreken echter voor de methode met shift.

TENTAMEN LINEAIRE ALGEBRA

(Numerieke Analyse II)

<u>ZATERDAG 8 MAART 1975</u>   9.30-12.30

- LEES ONDERSTAANDE INSTRUKTIES GOED DOOR
- ZET OP ELK INGELEVERD VEL UW NAAM
- ZET OP HET DUBBELE FOLIOVEL ALLEEN UW NAAM+ADRES

  DIT VEL MAG VERDER <u>NIET</u> BESCHREVEN WORDEN EN

  DIENT OM UW OPGAVEN IN TE LEVEREN
- U MAG VOORGAANDE ONDERDELEN GEBRUIKEN, OOK ALS

  ZE NIET BEWEZEN ZIJN.

1. We willen bewijzen dat voor alle reële n×n matrices A geldt:

$$\|A\|_2 = \|A\|_F \Leftrightarrow \text{rang}(A) \leq 1.$$

Zoals bekend bestaan er orthogonale matrices U en V en een niet negatieve diagonaalmatrix Σ zodat $A = U\Sigma V^T$.

Ga nu achtereenvolgens na:

(a) $\|\Sigma\|_2 = \|A\|_2$, $\|\Sigma\|_F = \|A\|_F$

(b) $\text{rang}(\Sigma) = \text{rang}(A)$ (bedenk dat $\text{rang}(A) = \dim(A\mathbb{R}^n)$)

(c) $\|A\|_2 = \|A\|_F \Leftrightarrow \text{rang}(A) \leq 1$.

2. Zij $A = \begin{pmatrix} 10 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$

(a) Geef afschattingen voor de eigenwaarden van A mbv. de stellingen van Gershgorin.

(b) Bepaal een diagonaal matrix D zodanig dat de stellingen van Gershgorin toegepast op DAD^{-1} aantonen dat er een eigenwaarde tussen 9.7 en 10.3 ligt.

(c) Laat zien dat 0 een eigenwaarde is van A, en toon vervolgens aan, zonder de karakteristieke veelterm op te stellen, dat de derde eigenwaarde tussen 1.7 en 2.3 ligt.

(d) Men kan A ook beschouwen als een symmetrisch geperturbeerde van diag (10, 1, 1). Dit geeft de mogelijkheid de eigenwaarden te schatten met behulp van de 2-norm van de stoormatrix (die hiertoe exact bepaald moet worden). Doe dit.

3. Zij A een $m\times m$ matrix, b een gegeven vector.

We wensen de oplossing van $Ax = b$ benaderd te bepalen.

Als norm kiezen we de supnorm.

(a) Zij $\tilde{x}$ een benadering van x. Bewijs nu dat $\|x-\tilde{x}\| \leqslant \|A^{-1}\|\|r(\tilde{x})\|$ met $r(\tilde{x}) = A\tilde{x}-b$. Ga ook na dat indien $\|A^{-1}\|\|r(\tilde{x})\| \leqslant \|\tilde{x}\|$ er geldt:

$$\frac{\|x-\tilde{x}\|}{\|x\|} \leqslant \frac{\|A^{-1}\|\|r(\tilde{x})\|}{\|\tilde{x}\|-\|A^{-1}\|\|r(\tilde{x})\|}\,.$$

(b) Stel nu dat x_0 de oplossing van het stelsel $Bx_0 = b$ is (B non-singulier!), waarbij $\|I-B^{-1}A\| < 1$ is.

Toon aan dat de rij vectoren $\{x_n\}$ met

(*) $x_n = (I-B^{-1}A)x_{n-1}+B^{-1}b$ $(n\geqslant 1)$ naar de exacte oplossing van $Ax = b$ convergeert (afgezien van afrond fouten).

(c) Zij B de matrix

$$\begin{bmatrix} 1 & 0\ldots 0 & \alpha_1 & & \\ & \ddots & \vdots & & \phi \\ & & \alpha_{k-1} & & \\ & & 1 & & \\ & & 1 & & \\ & \phi & \alpha_{k+1} & 1 & \\ & & \vdots & & \ddots \\ & & \alpha_m & & 1 \end{bmatrix}$$

(d) Geef een expliciete uitdrukking voor B^{-1} en toon aan dat $\|B^{-1}\| = 1+\max_{1\leqslant j\leqslant m}|\alpha_j|$.

(e) Zij nu $A = B+E$ met $\|E\|\leqslant \varepsilon$ en $\varepsilon(1+\max_{1\leqslant j\leqslant m}|\alpha_j|) < 0.1$.
Geef een zo scherp mogelijke afschatting voor $\|I-B^{-1}A\|$ en $\|A^{-1}\|$, en laat zien dat in elk geval $\|I-B^{-1}A\| < 0.1$.

(f) Stel $B^{-1}b$ is exact te bepalen; noteer $\|B^{-1}b\| = c$.

Toon aan $\frac{9}{10}c \leqslant \|x_n\| \leqslant \frac{10}{9}c$.

(g) Zij P een willekeurige computer representeerbare $m \times m$ matrix en x een dito vector. Laat zien dat de norm van de fout bij het berekenen van Px hoogstens $m\|P\|\|x\|\bar{\xi}$ is als $\bar{\xi}$ de max. relatieve machine afrondfout is en we hogere machten van $\bar{\xi}$ verwaarlozen.

(h) Neem nu aan dat de matrix $I-B^{-1}A$ en de vector $B^{-1}b$ exact bekend zijn en door de computer representeerbaar, en laat de computer de recursie (*) uitvoeren met deze exacte $I-B^{-1}A$. Opgeleverd wordt dan een rij $\{\tilde{x}_n\}$.

Toon aan

$$\|\tilde{x}_n-[(I-B^{-1}A)\tilde{x}_{n-1}+B^{-1}b]\| \leq [0.1(m+1)\|\tilde{x}_{n-1}\|+c]\bar{\xi},$$

als we hogere machten van $\bar{\xi}$ verwaarlozen.

(i) Zij $\delta_n = \tilde{x}_n-x_n$. Toon dan mbv. voorgaande resultaten en (*) aan

$$\|\delta_n\| \lesssim 0.1\|\delta_{n-1}\| + \frac{1}{9}(m+10)c\,\bar{\xi}$$

en laat hiermee zien dat $\|\delta_n\| \lesssim \frac{10}{81}(m+10)c\,\bar{\xi}$

(j) Toon nu aan dat voor n groot genoeg zeker geldt

$$\|\tilde{x}_n-x\| \lesssim \frac{10}{81}(m+10)c\,\bar{\xi}.$$